Submitted to the Electronic Journal of Statistics
Vol. 0 (2009)
ISSN: 1935-7524

Non-asymptotic model selection for linear non least-squares estimation in regression models and inverse problems

Ikhlef Bechar[*,†]

OCMR / FMRIB
Oxford University
John Radcliffe Hospital
Oxford, OX3 9DU
United Kingdom
e-mail: ikhlef.bechar@cardiov.ox.ac.uk

Projet Pulsar
INRIA Sophia Antipolis
Route des Lucioles - BP 93
06902 Sophia Antipolis Cedex
FRANCE
e-mail: ikhlef.bechar@sophia.inria.fr

Abstract: We propose to address the common problem of linear estimation in linear statistical models by using a model selection approach via penalization. Depending then on the framework in which the linear statistical model is considered namely the regression framework or the inverse problem framework, a data-driven model selection criterion is obtained either under general assumptions, or under the mild assumption of model identifiability respectively. The proposed approach was stimulated by the important recent non-asymptotic model selection results due to Birgé and Massart mainly (12), and our results in this paper, like theirs, are non-asymptotic and turn to be sharp.

Our main contribution in this paper resides in the fact that these linear estimators are not necessarily least-squares estimators but can be any linear estimators. The proposed approach finds therefore potential applications in countless fields of engineering and applied science (image science, signal processing, applied statistics, coding, to name a few) in which one is interested in recovering some unknown vector quantity of interest as the one, for example, which achieves the best trade-off between a term of fidelity to data, and a term of regularity or/and parsimony of the solution. The proposed approach provides then such applications with an interesting model selection framework that allows them to achieve such a goal.

AMS 2000 subject classifications: 62G08.
Keywords and phrases: Linear statistical model, Regression framework, Inverse problem framework, Regularity/parsimony priors, Non least-squares estimators, Non asymptotic Model selection, Penalized criterion, Oracle inequalities.

[*]EPSRC Postdoctoral Research Grant, Oxford University
[†]Pulsar project, INRIA Sophia Antiplois

Summary of the main results

Let us fix first some of the notations that we shall use in the sequel. So, we consider the problem estimation of a vector quantity $\beta = (\beta)_{k=1,p}$ lying in some $p-$dimensional Euclidean subspace of $\mathbf{R}^p$ (though our results extend easily to infinite dimensional Hilbert spaces) endowed with the traditional Euclidean norm $\|\cdot\|$ defined as follows

$$\|\mu\|^2 = \sum_{k=1}^{p} \mu_k^2,\ \forall \mu = (\mu_k)_{k=1,p} \in \mathbf{R}^p$$

We shall also use the same notation, i.e., $\|A\|$ to mean the Euclidean norm of any q by q real matrix A for some $q \in \mathbb{N}$, i.e.,

$$\|A\|^2 = \sum_{i=1}^{q}\sum_{j=1}^{q} a_{i,j}^2,\ \forall\, A = (a_{ij})_{i=1,q;j=1,q},\ a_{ij} \in \mathbf{R},\ \forall\, i,j = 1,\cdots,q$$

Sometimes, we use for some $q \in \mathbb{N}$ the notation I_q to mean the q by q identity matrix, and the notation $D\{\eta_k\}_{k=1,q}$ to mean a q by q diagonal matrix with diagonal elements $\eta_k, k = 1,\cdots,q$. When we write for some matrix A the following $A^\dagger$, we mean the Moore-Penrose pseudo-inverse of matrix A.

With this being said, we consider the classical linear gaussian regression model $y = X\beta + Rz$, and we assume that one is given a collection of linear estimators of the solution β as follows $\mathcal{L} = \{\hat{\beta}_m := \Psi_m y, m \in \mathcal{M}\}$ parameterized by some set of models (parameters) $\mathcal{M}$, that we consider finite in this paper though our results generalize easily to the case when $\mathcal{M}$ is countable-infinite. Such a collection of linear estimators can be obtained for instance by considering one (or more) of the frameworks presented in section 2. The goal is then to select among such a finite collection of linear estimators one estimator of β with the lowest quadratic risk (our results extend easily to other performance measures such as the weighted quadratic risk or the Mahalanobis risk). To do this, we firstly assume that when matrix X is rank-deficient, the solution of interest β is linearly-identifiable in the linear model above; which means that one knows a-priori some p by n real matrix $\mathcal{K}$ such that one can write $\beta - \mathcal{K}X\beta = 0$ [1], and we adopt a model selection procedure from a non-asymptotic point of view via penalization. Thus, we propose to select the estimator of β by minimizing over $\mathcal{M}$ a penalized criterion of the form

$$\mathrm{Crit}(m) = y^T\Big(\Psi_m^T\Psi_m - 2\mathcal{K}^T\Psi_m)y + \mathrm{pen}(m)$$

where $\mathrm{pen}(m), m \in \mathcal{M}$ stands for some penalty function that depends exclusively upon a given model m through matrix Ψ_m but not upon data that we propose to address in the remainder of this paper, and we refer to the estimator denoted by $\tilde{\beta} = \hat{\beta}_{\tilde{m}}$ with $\tilde{m} = \arg\inf_{m\in\mathcal{M}} \mathrm{Crit}(m)$ as the penalized estimator

[1] When X is of full rank, then one takes $\mathcal{K} := (X^TX)^{-1}X^T$.

of β. Then, we show that when the size of the collection of estimators $\mathcal{L}$ is not too great, one can set the penalty pen(m) in such a way to enforce an oracle inequality of the form

$$\mathbb{E}\Big[\|\tilde{\beta}-\beta\|^2\Big] \leq K \min_{m\in\mathcal{M}} \mathbb{E}\Big[\|\hat{\beta}_m-\beta\|^2\Big] + C$$

for some universal additive constant C and some multiplicative constant K which depends on the complexity of the model collection $\mathcal{M}$.
More generally, especially when the size of the collection $\mathcal{L}$ can be great, then we show that another choice of pen(m) that takes into account the complexity of each model with respect to the collection of models $\mathcal{M}$–in a sense that we shall specify in the sequel– warrants to obtain a sharp inequality of the form

$$\mathbb{E}\Big[\|\tilde{\beta}-\beta\|^2\Big] \leq K \min_{m\in\mathcal{M}} \Big\{\mathbb{E}\Big[\|\hat{\beta}_m-\beta\|^2\Big] + T_m\Big\} + C$$

with K and C standing for some reasonable universal constants, and T_m stands for some reasonable per-model positive quantity which is related to the complexity of a given model m with respect to the model collection $\mathcal{M}$.

1. Introduction

We assume the correlated linear statistical model

$$y = X\beta + Rz \tag{1.1}$$

where y is a n-dimensional vector of observations, β is the $p-$dimensional vector parameter of interest, X and R are n by p and n by n design (known) matrices respectively, and we take z to be a $p-$dimensional vector of independent and identically distributed (i.i.d.) random variables $N(0,1)$. We are then interested in estimating the vector β (resp. the response $X\beta$) given a single realization of the vector y by adopting a model selection approach via penalization.

We propose to address such an estimation problem under rather general assumptions about the model parameters X, R, n and p and the solution β. So, we regard β as an unknown deterministic vector quantity, and we assume that matrix R is a general n by n matrix, matrix X can be a well or an ill-conditioned matrix, the number of the observations (n) can be greater, equal or smaller than the size of the vector β (p), which therefore includes the case known as the "$n < p$" set up where on seeks to recover high resolution or sparse vector quantities β by using only a few measurements (18; 16).

Our results in this paper are non-asymptotic; this means that we propose to work with the values of the parameters R, n and p of the linear model as they are, and we allow the number of models of the solution β to depend freely upon the of the solution (p). This viewpoint as initiated in model selection by Barron, Birgé and Massart (7) then refined by Birgé and Massart (11; 12), needs to be contrasted with the asymptotic point of view (1; 23; 28; 29; 21) which considers for example that the number of the observations goes to infinity or that the

noise magnitude goes to zero while the number of models remains fix. As we shall see in the remainder, useful model selection criteria are directly connected to the complexity of a family of models, and by the latter, we roughly mean how large a model collection is; compared with the size (p) of the vector β [2]. Such a non-asymptotic property turns indeed to be very precious in many practical applications, mainly those which seek to recover a given vector quantity of interest by using a large library of models. We refer the interested reader to (12; 25) for a more thorough discussion on the topic.

Non-asymptotic model selection by using penalization has become an important trend in statistical estimation in regression, and will certainly continue in fascinating many researchers either in statistics or in the engineering field and applied science for a long time. Early works in the field appeared indeed in the early nineties due to Barron and colleague (6) for discrete models, and extensions to continuous models were proposed by Barron, Birgé and Massart (7). Aware of the pioneering works of Talagrand on concentration inequalities (31) [3], Birgé and Massart (11; 12) improved on Talgrand's works and refined the model selection approach in (7) and addressed nicely the problem of the estimation of the mean of a gaussian process in homoscedastic regression models when the variance is known (or estimated off-line), while taking into consideration the richness (complexity) of a collection of models, and which leads in some cases (when the model collection is not too big) to sharp oracle inequalities. Baraud (2; 4) and Baraud and colleagues (3) proposed many extensions to the aforementioned works that generalize the approach to non-gaussian homoscedastic statistical models; and under some mild assumptions about noise moments, amazingly they obtained near-oracle results. Baraud, Giraud and Huet (5) proposed recently penalized model selection criteria that are able to estimate the mean of gaussian homoscedastic models even when the variance is unknown, and they proved results for both the quadratic risk and the Kullbak risk. Very recently, Gendre (19; 20) extended their results in the case of a simultaneous estimation of the mean and the variance in gaussian heteroscedastic models. The latter work is probably the closest to our present work since it can handle heteroscedastic data as well, however, a subtle nuance exists between the two methods. Indeed while our method works with any linear estimators, the method of the author is based solely on least-squares estimators; which means that for each model m of the model collection (assumed to be some family of subspaces to which one can associate orthonormal bases), an estimator is constructed as the closest $p-$dimensional real vector to data (in the sense of the squared error) by assuming that current model m is true. Though we have to admit that such a work constitutes a pretty important contribution to the field

[2] It is interesting to note that we do not say with (n) because, anyway, p and n are related since they have to satisfy generally that $n \succeq O(\frac{S}{\log[p/n]})$ where S stands for the maximum number of the non-zero components of the vector β or of its coefficients with respect to some fixed orthonormal basis, otherwise it is not generally possible to recover efficiently β even when the latter is some high-resolution signal(14).

[3] We refer the dear reader to (24; 25) for two beautiful lectures on the topic of concentration inequalities and their application in model selection.

of model selection in correlated regression models, the approach of the author might suffer for yielding the expected results for some instances of the vector β and of the noise matrix R especially when β has many nonzero coefficients with respect to any basis among the family of bases of the solution, and/or matrix R plays a preponderant role in the formula of the quadratic risk. However, when such a *least-squares* restriction is relaxed to allow the construction of non least-squares linear estimators of the solution (which is the case here), interestingly, strong priors about the sought solution β can be inserted in a plenty of ways into its linear estimators (for instance, as the best trade-off between fidelity to data and regularity), consequently, one might limit considerably the influence of matrix R on the recovery process[4]. Interestingly, this finds countless applications in the fields of engineering and applied science–witness the broad success of Bayesian regularization frameworks in signal and image restoration. We would not close this rather modest overview of related works without probably pointing out one thriving field of non-asymptotic estimation in regression pioneered mainly by Candès and collaborators (13–16) and currently explored by many research groups in statistics and engineering disciplines (image and signal processing, information and coding theory, etc.) and which is known as sparse statistical estimation or compressive sensing. In the latter, one tries to recover some vector quantity of interest assumed to be sparse or admits a parsimonious representation in a fixed orthonormal basis by considering criteria based on the ℓ_1-norm minimization by using linear programming concepts. The authors obtained consequently under some minimal assumptions about the model near-oracle inequalities of their estimation method.

Before describing in details the main model selection results in this paper, we would like to open here a discussion to briefly point out our point of view regarding which performance measure of an estimator of β is better suited for some application; for sake of making our approach in this paper as much clear as possible, and for helping the dear reader set up more easily his/her model selection framework. In fact, one has to distinguish generally between the following two frameworks in which the linear statistical model (1.1) can be considered which we briefly review:

1. **The linear regression framework**: where the estimation of $X\beta$ is generally, though not always[5], the main concern for the statistician. Hence, the performance measure of any estimator $\hat{\beta}$ of β which is used in this case is the predictive risk given by $\mathbb{E}\left[\|X\hat{\beta} - X\beta\|^2\right]$.
2. **The linear inverse problem framework**: where the estimation of β is

[4] We would like then to point out that this paper is devoted to linear estimation by model selection in a broad context, and another paper (8) which deals specifically with the problem of construction of non least-squares estimators that can take into account to some extent data heteroscedasticity in the goal of achieving more interesting balances between the bias and the variance terms in the formula of the quadratic risk is currently in the writing process by the author in the spirit of the present work.

[5] because one would also use the regression framework to select an estimator of β, but it is to use with some care to avoid any bad situation of selecting an estimator $\hat{\beta}$ such that $\mathbb{E}\left[\|X\hat{\beta} - X\beta\|^2\right] \approx 0$, but $\mathbb{E}\left[\|\hat{\beta} - \beta\|^2\right] \gg 0$.

the main concern for the statistician. Hence, the performance measure of any estimator $\hat{\beta}$ of β which is used in this case is the quadratic risk given by $\mathbb{E}\left[\|\hat{\beta}-\beta\|^2\right]$.

However, one shows easily that w.l.g., even to consider a collection of linear estimators of $X\beta$ as follows

$$\{\widehat{X\beta}_m := X\hat{\beta}_m := X\Psi_m y, m \in \mathcal{M}\}$$

and use as a figure of merit of any estimator $\widehat{X\beta} := X\hat{\beta}$ of $X\beta$ its quadratic risk, which is simply the predictive risk of the estimator $\hat{\beta}$ of β, one then finds oneself in the presence of the model selection problem we stated earlier in this section. Consequently, we shall address in the remainder the two problems, namely the linear regression problem and the linear inverse problem, in a same framework. However, in contrast to the regression set up, model identifiability might be an issue to consider in the inverse problem set up to derive useful model selection procedures, so we shall also provide some useful ideas that help overcome such an issue.

Regarding now the appropriateness of either performance measure of any estimator of β, it turns out indeed that in many engineering domains, the vector quantity of interest that has direct application is β and not $X\beta$ (14). In this case, using the quadratic risk as a figure of merit of an estimator of β makes more sense than using the predictive risk. In image reconstruction/restoration applications for instance, β represents an image (e.g. an MRI scan of the heart or the brain of a patient), X then models the design matrix of the imaging device (e.g. the MRI scanner), and Rz models the stochastic errors of the imaging device. Obviously, the vector quantity of interest in this case is the image β that one would need to reconstruct for further use (for cardiac or brain diagnosis for example). However, when one is more interested in the estimation of the system response $X\beta$ than β itself, then using the predictive risk as a figure of merit of an estimator sounds more interesting in this case. As some illustrative examples, one can mention the variable selection problem that we shall discuss in section 2, and the reconstruction problem (for filtering or recognition purposes, etc.) by using a finite library of vector primitives. In the latter, one has generally an a-priori linear model of some vector quantity of interest u (a signal, an image, a curve, a shape, etc.) as follows $u = X\beta$, where $\left\{X_k, k = 1, \cdots, p\right\}$ stand for some set of vector primitives (predictors) which, depending on the application, can be for example eigen objects (e.g. eigen images, eigen shapes, wavelets) or simply some database of generic objects, and β stands for the vector of the coefficients of the linear combination of such vector primitives. One's goal is then to estimate the vector of the coefficients β in such a way to achieve the lowest predictive error. To cut a long story short, depending on the application, one may find good reasons to prefer the use of one performance measure from the use of its alternative (see for example subsection 3.2 for another reason that might justify the use of the predictive risk).

Having said this, the rest of the paper is organized as follows. In section 2, we set up our model selection framework and we describe some applications that

can be expressed in terms of our framework. In section 3, we derive the main data-driven model selection results in this paper, and we provide some useful clues that help choose the penalty function and its parameters. In section 4, we discuss some practical solutions that can lead to overcome the identifiability issue of model (1.1) when the rank of matrix X is smaller than p. Section 5 is devoted to the numerical experiments that show the performances of the approach for some application examples. Finally, a general discussion about the proposed approach and its future extensions concludes this papers.

2. Collection of linear estimators of β

We assume that one is given typically a finite collection of linear estimators of the solution β which is parameterized by some finite set of models (or parameters) $\mathcal{M}$ as follows

$$\mathcal{L} = \{\hat{\beta}_m := \Psi_m y, m \in \mathcal{M}\} \tag{2.1}$$

with Ψ_m standing for some p by n matrix for all $m \in \mathcal{M}$, and the goal is to select among such a family of linear estimators $\left\{\hat{\beta}_m, m \in \mathcal{M}\right\}$ one estimator of β with the lowest quadratic risk.

Concerning the way in which these linear estimators are constructed, this depends generally on the application and on the prior that one has about the solution β, and some examples of their construction encountered in countless engineering domains (signal/image processing, computer vision, applied statistics, to name a few) are highlighted below.

2.1. Reconstruction/Restoration by regularization

In many image and signal reconstruction/restoration applications, for recovering some vector quantity of interest namely β from an observation of model (1.1), one proceeds by solving an unconstrained quadratic problem of the form

$$\left(y - X\hat{\beta}\right)^T P\left(y - X\hat{\beta}\right) + \hat{\beta}^T H\hat{\beta} \rightarrow \min_{\hat{\beta}\in\mathbf{R}^p} \tag{2.2}$$

where P and H stand for two given n by n and p by p symmetric matrices respectively, both considered generally to be positive semi-definite. The two competing terms $\left(y-X\hat{\beta}\right)^T P\left(y-X\hat{\beta}\right)$ and $\hat{\beta}^T H\hat{\beta}$ in (2.2) in (2.2) are generally referred as the data fidelity term, and the regularity term respectively. As an illustrative example, when β is assumed to be some smooth vector quantity, then H is taken generally to be some regularization operator (e.g. a high-pass filter such as a differential operator or any linear combination of differential operators of different orders, a band-pass filter, etc.) in order to enforce smoothness of the recovered solution, and P is generally taken as the inverse of the covariance matrix of the observations, i.e., $P = \left(R^T R\right)^{-1}$ or simply as the n by n identity matrix, i.e., $P = I_n$. This is also known in applied science as the Tikhonov

regularization problem (33) and which has Bayesian interpretation (maximum of posterior log-likelihood of data in a gaussian set up). Solving now for $\hat{\beta}$ in (2.2) gives

$$\hat{\beta} = \left(X^T P X + H\right)^{\dagger} X^T P y$$

However, it is quite common that matrix H and probably matrix P too (useful if one wants for example to test the performance of the linear estimators against various Mahalanobis distances between data and an estimator's response) depend on some parameters (referred in engineering as regularization parameters) that are difficult to tune optimally for a particular application. Therefore, one assumes a finite collection of parametric couples of instances of the matrices P and H as follows $\{\left(P_m, H_m\right), m \in \mathcal{M}\}$, and by putting for all $m \in \mathcal{M}$

$$\Psi_m := \left(X^T P_m X + H_m\right)^{\dagger} X^T P_m$$

one obtains finally a finite parametric collection of linear estimators of β of the form $\left\{\hat{\beta}_m := \Psi_m y, m \in \mathcal{M}\right\}$. The goal is then to select among such a parametric collection of linear estimators of β one estimator with the lowest risk. One then finds oneself in the presence of the model selection framework that we set up in the beginning of this section.

2.2. Optimal representation in a family of orthonormal bases

Some applications that attempt to estimate β from an observation of model (1.1) assume that β has a sparse representation in a given family of orthonormal bases in $\mathbf{R}^p$ of different complexities (i.e., dimensions) $\Lambda = \{\Lambda_m, m \in \mathcal{M}\}$ (e.g. trigonometric bases, a wavelet family, etc.), and let us denote by $\overline{\Phi}_m$ the matrix which rows correspond to the respective vectors of the complement basis in $\mathbf{R}^p$ of the orthonormal basis Λ_m, for all $m \in \mathcal{M}$. So, these applications proceed indeed by solving for all $m \in \mathcal{M}$ a constrained quadratic program of the form

$$\left(y - X\hat{\beta}_m\right)^T P_m \left(y - X\hat{\beta}_m\right) \rightarrow \min_{\hat{\beta}_m / \overline{\Phi}_m \hat{\beta}_m = 0} \tag{2.3}$$

where P_m stands for a $n-$dimensional symmetric matrix. Let us put

$$C_m = (X^T P_m X + \overline{\Phi}_m^T \overline{\Phi}_m)^{\dagger}$$

One checks that the optimal estimator $\hat{\beta}_m^*$ with respect to $m \in \mathcal{M}$ is given by

$$\hat{\beta}_m = C_m \left[I_p - \overline{\Phi}_m^T \left(\overline{\Phi}_m C_m \overline{\Phi}_m^T\right)^{\dagger} \overline{\Phi}_m C_m\right] X^T P_m y$$

moreover, if matrix $\left(X^T P_m X + \overline{\Phi}_m^T \overline{\Phi}_m\right)$ is of full rank (i.e., p), then such an estimator is the unique solution of (2.3). Let us now put for all $m \in \mathcal{M}$

$$\Psi_m := C_m \left[I_p - \overline{\Phi}_m^T \left(\overline{\Phi}_m C_m \overline{\Phi}_m^T\right)^{\dagger} \overline{\Phi}_m C_m\right] X^T P_m$$

hence one obtains a finite collection of linear estimators of β which is given by $\left\{\hat{\beta}_m := \Psi_m y, m \in \mathcal{M}\right\}$, among which one wants to select one estimator with the lowest risk. One retrieves again our model selection framework.

Remark 1. *The formula of the solution $\hat{\beta}_m$ of (2.3) and all those that shall come up later in this section remain valid for arbitrary matrices P_m and $\overline{\Phi}_m$, and it is not definitely necessary to assume that the row vectors of $\overline{\Phi}_m$ are orthonormal (hence when Φ_m is orthogonal, one may replace $\overline{\Phi}_m$ with $I_p - \Phi_m^T\Phi_m$). Please note that one can approximate the formula of $\hat{\beta}_m$ by a simpler formula for some large enough positive number μ as follows $\hat{\beta}_m \approx (X^T P_m X + \mu\overline{\Phi}_m^T\overline{\Phi}_m)^\dagger X^T P_m y$.*

2.3. Optimal representation in a family of orthonormal bases with regularization

It happens that the solution β that one looks to estimate from an observation of model (1.1) has a sparse representation in some family Λ of orthonormal bases of some subspaces in $\mathbf{R}^p$ which we write as $\Lambda = \left\{\Lambda_{m'}, m' \in \mathcal{M}^{orth}\right\}$ and has some regular (e.g. smooth) structure. So, let us denote for all $m' \in \mathcal{M}^{orth}$ by $\Phi_{m'}$ the matrix which rows correspond to the respective vectors of the orthonormal basis $\Lambda_{m'}$, and by $\overline{\Phi}_{m'}$ the matrix which rows correspond to the respective vectors of the complement basis $\overline{\Lambda}_{m'}$ in $\mathbf{R}^p$ of $\Lambda_{m'}$ and put $\overline{\Lambda} = \left\{\overline{\Lambda}_{m'}, m' \in \mathcal{M}^{orth}\right\}$. The goal is then to recover the solution of β in the orthonormal family Λ as parsimoniously as possible, while imposing either on β or on the vector of coefficients of β in any orthonormal basis $\Lambda_{m'} \in \Lambda$ to have a regular profile (e.g. smoothness), but the latter is generally difficult to guess beforehand. Hence, one would like to consider on top of Λ a finite parametric collection $\mathcal{H}$ of regularizing operators with increasing regularization powers as follows $\mathcal{H} = \left\{H_\theta, \theta \in \Theta\right\}$, and depending on whether the smoothing is applied on β directly or on its vector of coefficients with respect to any orthonormal basis $\Lambda_{m'} \in \Lambda$, an operator $H_\theta \in \mathcal{H}$ may depend or not on a basis $\Lambda_{m'}$. Nonetheless, for the sake of simplicity and without loss of generality (w.l.g.) (see remark 2 below), one can consider that for all $\theta \in \Theta$, H_θ is p by p symmetric matrix and that one is given a family of models (quadruplets) of the form $\left\{(\Lambda_m, \overline{\Lambda}_m, H_m, P_m), m \in \mathcal{M}\right\}$ where for all $m \in \mathcal{M}$, $\Lambda_m \in \Lambda$, $\overline{\Lambda}_m \in \overline{\Lambda}$, $H_m \in \mathcal{H}$, and P_m is some n by n symmetric matrix which enforces closeness of an estimator to data. Then with respect to all $m \in \mathcal{M}$, one solves a constrained quadratic program of the form

$$\left(y - X\hat{\beta}\right)^T P_m \left(y - X\hat{\beta}\right) + \beta^T H_m \beta \rightarrow \min_{\hat{\beta}/\overline{\Phi}_m\hat{\beta}=0}$$

Let us put for all $m \in \mathcal{M}$

$$C_m = (X^T P_m X + \overline{\Phi}_m^T\overline{\Phi}_m + H_m)^\dagger$$

One checks (see the proof in the appendix section) that the optimal estimator $\hat{\beta}_m^*$ with respect to $m \in \mathcal{M}$ is given by

$$\hat{\beta}_m = C_m \left[I_p - \overline{\Phi}_m^T \left(\overline{\Phi}_m C_m \overline{\Phi}_m^T \right)^\dagger \overline{\Phi}_m C_m \right] X^T P_m y$$

and one can approximate the latter for a large enough positive number μ as follows

$$\hat{\beta}_m \approx (X^T P_m X + H_m + \mu \overline{\Phi}_m^T \overline{\Phi}_m)^\dagger X^T P_m y$$

Let us now put for all $m \in \mathcal{M}$

$$\Psi_m := C_m \left[I_p - \overline{\Phi}_m^T \left(\overline{\Phi}_m C_m \overline{\Phi}_m^T \right)^\dagger \overline{\Phi}_m C_m \right] X^T P_m$$

One obtains in the end the following finite collection of linear estimators of β: $\left\{ \hat{\beta}_m := \Psi_m y, m \in \mathcal{M} \right\}$ among which one would want to select one estimator with the lowest risk. One retrieves again our model selection framework.

Remark 2. *When the smoothness constraint is imposed rather on the coefficients of an estimator $\hat{\beta}_m$ in a given orthonormal basis $\Lambda_m \in \Lambda$, one can proceed in the same way as above by defining for all $m \in \mathcal{M}$ the new filter H_m as follows $H_m := \Phi_m^T F_m \Phi_m$ with F_m standing for the filter which operates on the vector of the coefficients of $\hat{\beta}_m$ with respect to Λ_m, and H_m standing for the new filter that operates directly on the reconstructed solution $\hat{\beta}_m$.*

2.4. Optimal linear filtering

Linear filtering (low-pass, high-pass, band-pass, band-stop, etc.) is an important pre-processing task in various signal and image processing applications. For instance, low pass filtering (e.g. gaussian filtering) has become a standard step of the image processing chain which aims at removing white noise from a noisy image y in order to improve its quality (measured generally as peak signal-to-noise ratio (PSNR)) and simplify its further analysis and interpretation. However, linear filters depend generally on some parameters (e.g. a filter's bandwidth) which are difficult to tune optimally for a particular signal or an image. Hence one would need to consider a finite collection of parametric linear filters as follows $\{\Psi_m, m \in \mathcal{M}\}$ from which one would like to select the one with the best parameter to apply on the noisy image/signal y. Such a problem can be formulated indeed in terms of our model selection framework by considering a finite collection of linear estimators of the noiseless signal or image β as follows $\{\hat{\beta}_m := \Psi_m y, m \in \mathcal{M}\}$, therefore the goal amounts to selecting one estimator of β with the lowest quadratic risk.

2.5. Variable selection in regression

In this problem, one commonly assumes a collection of possible configurations $\mathcal{C} = \left\{ \nu_m, m \in \mathcal{M} \right\}$ of β, where each configuration ν stands for a binary vector

of the same size as β, and a component $\nu_k = 1$ means that the predictor X_k (i.e., the $k-th$ column of matrix X) is not part of the regression model (i.e., $\nu_k = 1$, $\beta_k = 0$), otherwise it is said to be part of the regression model (i.e., $\nu_k = 0$, $\beta_k \neq 0$), and the goal is to figure out the configuration of β that achieves lowest predictive error (see below for more insight on the use of the predictive risk), in other words, the most influential variables (i.e., those with most explanatory power) in model (1.1) among the set of variables $\{X_k, k = 1, \cdots, p\}$. So, let us denote for all $m \in \mathcal{M}$ by N_m the $p-$dimensional diagonal matrix such that $N_m(k,k) = \nu_k, k = 1, \cdots, p$. Hence one proceeds by minimizing with respect to each configuration $\nu_m \in \mathcal{C}$ a constrained quadratic program of the form

$$\left(y - X\hat{\beta}_m\right)^T P_m \left(y - X\hat{\beta}_m\right) \rightarrow \min_{\hat{\beta}/N_m\hat{\beta}=0}$$

for some $n-$dimensional symmetric matrix P_m to achieve finally a collection of linear estimators of β as follows: $\left\{\hat{\beta}_m := \Psi_m y, m \in \mathcal{M}\right\}$ where for all $m \in \mathcal{M}$, one has Ψ_m which is given by

$$\hat{\beta}_m = C_m \left[I_p - N_m \left(N_m C_m N_m\right)^{\dagger} N_m C_m\right] X^T P_m y$$

with

$$C_m = (X^T P_m X + N_m)^{\dagger}$$

We would like to emphasize that our goal in the remainder is not to address the problem of construction of the matrices P_m, Φ_m, H_m, N_m for a given application since they are application-dependent and they are constructed from the a-priori knowledge about the solution as we already mentioned it above [6]. Nevertheless, our goal–once a finite collection of them has been chosen–is to provide the user with an efficient model selection tool that allows him/her to choose the almost best ones to recover the solution.

Before starting in addressing such a model selection issue, we would like to enunciate upfront the following proposition which gives the exact formula of the ideal linear estimator of β with respect to the quadratic risk (see formula (3.2) in section 3).

Proposition 1. *The optimal linear estimator $\hat{\beta}^*$ that minimizes in $\mathbf{R}^p$ the formula of the quadratic risk is given by*

$$\hat{\beta}^* = \beta(X\beta)^T \left((X\beta)(X\beta)^T + RR^T\right)^{\dagger} y$$

[6] in fact, the experts in these domains have already done a great job in this respect, and a overview of the subject in this paper would only be a weak copy of their previous findings ...

and its quadratic risk is given by

$$\mathbb{E}\Big[\|\hat{\beta}^* - \beta\|^2\Big] = \|\beta\|^2\Big[$$
$$(X\beta)^T\Big((X\beta)(X\beta)^T + RR^T\Big)^{\dagger^T}(X\beta)(X\beta)^T\Big((X\beta)(X\beta)^T + RR^T\Big)^{\dagger}(X\beta)$$
$$-2(X\beta)^T\Big((X\beta)(X\beta)^T + RR^T\Big)^{\dagger}(X\beta)+\Big\|\Big((X\beta)(X\beta)^T + RR^T\Big)^{\dagger}(X\beta)R\Big\|^2+1\Big]$$

Proof. The proof of this proposition consists of a simple minimization of formula (3.3) of section 4 which thus amounts to solving an unconstrained quadratic program. □

Remark 3. *As an aside, it is interesting to note that proposition 1 says among other things that the solution with the lowest risk of model (1.1) corresponds to the least-norm solution of* $y' = X\beta$*, which is nothing else than* $X^{\dagger}y' \equiv X^{\dagger}X\beta$*. However, the latter does not correspond generally to the solution that one is looking to estimate, hence the estimation of the actual solution would have unavoidably a bigger risk.*

Though such a result of proposition (1) is of little use in practice since it says that such an ideal filter depends on the unknown β itself, nevertheless, it might provide some useful hints for the construction of potential linear models of the solution for some applications. Moreover, when the object β that one would like to estimate represents some object among a finite library of objects, then spanning such a library once allows to construct potential linear estimators of the solution among which the ideal one exists. For instance, in recognition or tracking tasks in computer vision, the unknown β might model a given physical object of interest that was previously extracted from some image for example by using any segmentation method, and one wants actually to recognize it among a finite library of objects $\mathcal{O} = \{O_m, m \in \mathcal{M}\}$. Then to each object $O_m, m \in \mathcal{M}$, one associates a $p-$dimensional vector of features β_m (for instance, some discretization of the contour of O_m), matrix X stands for some linear geometric transformation (rigid, affine, projective, etc.) applied onto the object, $X\beta$ represents the object β after deformation and Rz represents noise (and probably the errors due to the linear approximation of the deformation). Hence, one might consider the finite collection of linear estimators of β as follows: $\{\hat{\beta}_m = \Psi_m y, m \in \mathcal{M}\}$ with $\Psi_m := \beta_m(X\beta_m)^T\Big((X\beta_m)(X\beta_m)^T + RR^T\Big)^{\dagger}$ for all $m \in \mathcal{M}$, and use the model selection procedure that we present in the sequel to select the best one i.e., to recognize the actual object.

We should point out that the same approach might be used also in estimation; actually when one has a-priori a finite collection of profiles of the vector of the coefficients of the solution with respect to some orthonormal basis (or a family of orthonormal bases). These profiles may be obtained, for example, if one knows that the solution belongs to some class of $p-$dimensional vectors (assumed to be some discretization of continuous functions) such that the vector of the coefficients of each vector u of this class with respect to some p by p

orthogonal matrix $\Phi_{m'}$ can be bounded sharply by some parametric function (a $p-$dimensional vector actually) $f_{m'}(\mu(u))$ which parameter $\mu(u)$ should be of much lower dimension than p of course (typically 2 or 3). In this case, one may proceed by instantiating a finite collection of instances of the parameter $\mu(u)$ as follows $\{\mu_m, m \in \mathcal{M}\}$, and constructs a finite family of linear estimators of β as follows $\{\hat{\beta}_m := \Psi_m, m \in \mathcal{M}\}$, with $\Psi_m := \beta_m(X\beta_m)^T\Big((X\beta_m)(X\beta_m)^T + RR^T\Big)^\dagger$ and $\beta_m = \Phi_{m'}^T f(\mu_m)$ for all $m \in \mathcal{M}$.

3. Data-driven model selection for linear regression and linear inverse problems

Keeping in mind the linear regression model (1.1) and the collection of linear estimators of β introduced in (2.1), we suggest to devise a data-driven model selection criterion that can select a good estimator of β among such a collection of its linear estimators, and we measure the performance of any estimator $\hat{\beta}$ of β with its quadratic risk given by $\mathbb{E}\big[\|\hat{\beta} - \beta\|^2\big]$.

Clearly, when matrix X is rank deficient, i.e., its rank is inferior than p, model (1.1) is not identifiable. In this case, one needs to put some assumptions on the solution β to guarantee identifiability of the model and to make of its statistical inversion a possible question.Therefore, our identifiability hypothesis in this body of work consists in saying that one knows a-priori some linear operator $\mathcal{K}$–called a noiseless reconstructor of β– such that one can uniquely recover the unknown vector quantity of interest β from one noiseless observation of model (1.1) (i.e., if matrix R in (1.1) was zero) by using a linear formula of the form

$$\beta = \mathcal{K}X\beta \tag{3.1}$$

This shall be referred as the *linear identifiability condition*. The motivations for imposing an identifiability condition of type (3.1) on model (1.1) are presented in details in section 4. Please note that without any prior on β, one can still write $\beta = \mathcal{K}X\beta + \overline{\mathcal{K}}\beta$, with $\mathcal{K} = X^\dagger$ and $\overline{\mathcal{K}} = I_p - X^\dagger X$. In this case, only the component of β that belongs to the subspace $\mathcal{S} = \Big\{X^\dagger X t, t \in \mathbf{R}^p\Big\}$, in other words $\mu = X^\dagger X\beta$ (which is also the solution of $\|\mu\|^2 \to \min_{\mu\in\mathcal{S}}$) could be theoretically recovered, and the left part of β i.e., $\bar{\mu} = (I_p - X^\dagger X)\beta$ which belongs to the subspace $\mathbf{R}^p - \mathcal{S}$ hence is lost. Though it is generally of little use in practice, one would recover such a least-norm estimate of the solution of (4.1) by writing the new model $y = X\mu + Rz$, and one has μ which obeys the linear identifiability condition with $\mathcal{K} = X^\dagger$ since one can write $\mu = X^\dagger X\mu$ (see also section 4 for more details). In section 4, we shall describe some situations which can lead to derive useful expressions of the noiseless reconstructor $\mathcal{K}$ of β to allow recovery of solutions other than such a least-norm solution.

With this being said, let us now focus on the model selection problem we stated previously in section 2, and let us start by deriving for all $m \in \mathcal{M}$ the expression of the quadratic risk of an estimator $\hat{\beta}_m$ of β, for all $m \in \mathcal{M}$. So, one

fixes some $m \in \mathcal{M}$, and from (1.1) and (2.1) one has

$$\hat{\beta}_m = \Psi_m y = \Psi_m X\beta + \Psi_m Rz$$

hence

$$\hat{\beta}_m - \beta = \big(\Psi_m X - I\big)\beta + \Psi_m Rz$$

then, by rising both sides of the latter equality to power of two, one obtains

$$\|\hat{\beta}_m - \beta\|^2 = \beta^T\big(\Psi_m X - I_p\big)^T\big(\Psi_m X - I_p\big)\beta + z^T R^T \Psi_m^T \Psi_m Rz + 2\beta^T\big(\Psi_m X - I\big)^T \Psi_m Rz$$

Finally, by applying the expectation operator on both sides of the latter equality, one derives the following expression of the quadratic risk of $\hat{\beta}_m$

$$\mathbb{E}\Big[\|\hat{\beta}_m - \beta\|^2\Big] = \beta^T\big(\Psi_m X - I_p\big)^T\big(\Psi_m X - I_p\big)\beta + tr\Big(R^T \Psi_m^T \Psi_m R\Big)$$

or equivalently

$$\mathbb{E}\Big[\|\hat{\beta}_m - \beta\|^2\Big] = \|\beta\|^2 + \beta^T X^T \Psi_m^T \Psi_m X\beta - 2\beta^T \Psi_m X\beta + \|\Psi_m R\|^2 \qquad (3.2)$$

Formula (3.2) says that the quadratic risk of $\hat{\beta}_m$ decomposes as the sum of two terms; a bias term:

$$\|\beta\|^2 + \beta^T X^T \Psi_m^T \Psi_m X\beta - 2\beta^T \Psi_m X\beta$$

and a variance term:

$$\|\Psi_m R\|^2$$

and the best estimator of β denoted by $\hat{\beta}_{m*}$ for some $m^* \in \mathcal{M}$ is definitely the one that achieves the best bias-variance tradeoff in the formula of the quadratic risk (3.2); or equivalently (since $\|\beta\|^2$ is a constant), the one that minimizes over $\mathcal{M}$ the expression

$$\beta^T X^T \Psi_m^T \Psi_m X\beta - 2\beta^T \Psi_m X\beta + \|\Psi_m R\|^2 \qquad (3.3)$$

Unfortunately, since expression (3.3) that one ideally would like to minimize over $\mathcal{M}$ depends on the unknown β itself, hence such a minimization cannot be carried out in practice. Thus, we refer to the optimal but the unaccessible procedure that minimizes (3.3) over $\mathcal{M}$ as an oracle, and the latter shall serve as a benchmark for assessing an estimator's performance.

To the impossibility of minimizing exactly the expression of the quadratic risk adds, unfortunately, another more challenging difficulty which has to do directly with identifiability of model (1.1)[7]. Indeed, when the rank of matrix X is inferior than the size of the solution β (p), only the linear transformation $X\beta$ of β can

[7]Obviously, as explained earlier in this paper, when $X\beta$ is the quantity which one looks to estimate (i.e., in the regression case), model identifiability is not generally a concern.

be observed (up to noise), and because of the term $\beta^T\Psi_m X\beta$ in (3.3) which doesn't depends upon β only through $X\beta$, this makes it quite challenging to derive a satisfactory data-driven model selection procedure. Nonetheless, with the identifiability condition that we introduced earlier in this section, such a model selection procedure becomes possible. Indeed, assume that one can write $\beta = \mathcal{K}X\beta$, one deduces the following new formula of the quadratic risk

$$\mathbb{E}\Big[\|\hat{\beta}_m - \beta\|^2\Big] = \|\beta\|^2 + (X\beta)^T\Psi_m^T\Psi_m X\beta - 2(X\beta)^T\mathcal{K}^T\Psi_m X\beta + \|\Psi_m R\|^2$$

hence formula (3.3) that one would like to minimize over $\mathcal{M}$ becomes

$$(X\beta)^T\Big(\Psi_m^T\Psi_m - 2\mathcal{K}^T\Psi_m\Big)X\beta + \|\Psi_m R\|^2$$

The latter formula has the advantage to depend on β only through $X\beta$ which, as we shall see in the remainder, makes it now possible to derive useful data-driven model selection procedures.

3.1. Model selection

Now, we propose to select an estimator of β among its finite set of linear estimators (2.1) by minimizing a penalized criterion of the form

$$\mathrm{Crit}(m) = y^T\Big(\Psi_m^T\Psi_m - 2\mathcal{K}^T\Psi_m)y + \mathrm{pen}(m) \tag{3.4}$$

where $\mathrm{pen}(m)$ stands for an arbitrary penalty function which good choice shall be addressed later in this section. The estimator of β denoted by $\tilde{\beta} = \hat{\beta}_{\tilde{m}}$ that minimizes criterion (3.4) shall be referred as the penalized estimator of β, and its quadratic risk $\mathbb{E}\Big[\|\tilde{\beta} - \beta\|^2\Big]$ shall be referred as the penalized risk. The following theorem provides then some useful clues that help choose the penalty $\mathrm{pen}(m)$ in (3.4).

Theorem 1. *Assume model (1.1) along with the collection of linear estimators of β introduced in (2.1), and assume equality (3.1). Let us fix some positive number $\theta \in (0,1)$, and consider for all $m \in \mathcal{M}$ the matrices*

$$\mathcal{A}_m = R^T\Big(-\theta\Psi_m^T\Psi_m + \mathcal{K}^T\Psi_m + \Psi_m^T\mathcal{K}\Big)R$$

$$\mathcal{B}_m = \Big(\theta\Psi_m - \mathcal{K}\Big)RR^T\Big(\theta\Psi_m - \mathcal{K}\Big)^T$$

and denote by s_m^+ the largest positive eigen value of the symmetric matrix $\mathcal{A}_m$ and by r_m^* the largest singular value of matrix $\mathcal{B}_m$. Let us now associate to each model $m \in \mathcal{M}$ two real positive numbers L_m and h_m with the L_m's satisfying $\Sigma = \sum_{m\in\mathcal{M}}\exp[-L_m] < \infty$, and consider for all $m \in \mathcal{M}$ the quantity Q_m

defined as follows

$$Q_m = 2tr\Big(R^T\mathcal{K}^T\Psi_m R\Big) - \theta\|\Psi_m R\|^2 + 2\big(\frac{r_m^*}{\theta} + s_m^+\big)L_m + 2\|A_m\|\sqrt{L_m + h_m} + \lambda\Big(\frac{\|A_m\|}{\sqrt{L_m} + \sqrt{L_m + h_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)^2$$

for some positive number λ. It follows that for every penalty function $pen(m), m \in \mathcal{M}$, the corresponding penalized estimator $\tilde{\beta}$ satisfies

$$(1-\theta)\mathbb{E}\Big[\|\tilde{\beta}-\beta\|^2\Big] \leq \inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2 + \beta^T X^T\Psi_m^T\Psi_m X\beta - 2\beta^T\Psi_m X\beta + \|\Psi_m R\|^2 - 2tr\Big(R^T\mathcal{K}^T\Psi_m R\Big) + pen(m)\Big\} + \sup_{m\in\mathcal{M}}\Big(Q_m - pen(m)\Big) + \frac{2\Sigma}{\lambda}$$

The proof of this theorem is deferred to the appendix section.

3.2. Equivalent of the model selection result in the regression framework

Though we considered the inverse problem framework and we argued that this was w.l.g., however before commenting in details theorem 1, we would like to give its equivalent in the regression framework as this might help make our approach as much clear as possible. Another motivation that we would like to emphasize and which we had already the opportunity to discuss in the beginning of this section is that, in contrast to the inverse problem framework, the identifiability assumption about model (1.1) is not necessary in the regression framework to use the proposed model selection approach. Indeed, such a linear identifiability condition might turn to be difficult to establish in some applications, however, one might use instead some priors about the solution (see the examples in section 2) so as to enhance identification of the solution, and feel rather safe to use the predictive risk as an estimator's measure of performance.

Now, consider the finite collection of linear estimators of β given in (2.1), then one establishes the following formula of the predictive risk of $\hat{\beta}_m$, for all $m \in \mathcal{M}$

$$\mathbb{E}\Big[X\|\hat{\beta}_m - X\beta_m\|^2\Big] = \|X\beta\|^2 + (X\beta)^T\Big(\Psi_m^T X^T X\Psi_m - 2X\Psi_m\Big)(X\beta) + \|X\Psi_m R\|^2$$

and one selects the model of β by minimizing over $\mathcal{M}$ a penalized criterion of the form

$$\text{Crit}_{pred}(m) = y^T\big(\Psi_m^T X^T X\Psi_m - 2X\Psi_m\big)y + \text{pen}_{pred}(m) \tag{3.5}$$

for an arbitrary penalty $\text{pen}^{pred}(m), m \in \mathcal{M}$. Let us denote by $\tilde{\beta}^{pred} = \hat{\beta}_{\tilde{m}}$ the estimator of β that minimizes (3.5) over $\mathcal{M}$, then the following theorem provides an upper bound of the predictive risk of $\tilde{\beta}^{pred}$.

Theorem 2. *Assume the linear model (1.1), along with the finite collection linear estimators of β introduced in (2.1). Let us fix some positive number $\theta \in (0,1)$, and consider for all $m \in \mathcal{M}$ the matrices*

$$\mathcal{A}_m^{pred} = R^T\big(-\theta\Psi_m^T X^T X\Psi_m + X\Psi_m + \Psi_m^T X^T\big)R$$

$$\mathcal{B}_m^{pred} = \big(\theta X\Psi_m - I_n\big)RR^T\big(\theta X\Psi_m - I_n\big)^T$$

and denote by s_m^+ the largest positive eigen value of the symmetric matrix $\mathcal{A}_m^{pred}$ and by r_m^* the largest singular value of matrix $\mathcal{B}_m^{pred}$. Let us now associate to each model $m \in \mathcal{M}$ two real positive numbers L_m^{pred} and h_m^{pred} with the L_m^{pred}'s satisfying $\Sigma^{pred} = \sum_{m\in\mathcal{M}} \exp[-L_m^{pred}] < \infty$, and consider for $m \in \mathcal{M}$ the quantity

$$Q_m^{pred} = 2tr\Big(R^T X\Psi_m R\Big) - \theta\|X\Psi_m R\|^2 + 2\Big(\frac{r_m^*}{\theta} + s_m^+\Big)L_m^{pred} + 2\|A_m\|\sqrt{L_m^{pred} + h_m^{pred}} + \lambda\Big(\frac{\|A_m\|}{\sqrt{L_m^{pred}} + \sqrt{L_m^{pred} + h_m^{pred}}} + \frac{r_m^*}{\theta} + s_m^+\Big)^2$$

for some positive number λ. It then follows that for every penalty function $pen_{pred}(m)$ for all $m \in \mathcal{M}$, the selected estimator $\tilde{\beta}^{pred}$ satisfies

$$(1-\theta)\mathbb{E}\Big[\|X\tilde{\beta}^{pred} - X\beta\|^2\Big] \le \inf_{m\in\mathcal{M}}\Big\{\|X\beta\|^2 + (X\beta)^T\Big(\Psi_m^T X^T X\Psi_m - 2X\Psi_m\Big)(X\beta) + \|X\Psi_m R\|^2 - 2tr\Big(R^T X\Psi_m R\Big) + pen_{pred}(m)\Big\} + \sup_{m\in\mathcal{M}}\Big\{Q_m^{pred} - pen_{pred}(m)\Big\} + \frac{2\Sigma^{pred}}{\lambda}$$

Proof. Theorem 2 is a simple consequence of theorem 1 by taking into account the remarks above about the fact that considering the quadratic risk as a performance measure of an estimator $\widehat{X\beta}$ of $X\beta$ amounts to considering the predictive risk as a performance measure of an estimator $\hat{\beta}$ of β. □

As such a model selection result in the case of the predictive risk is a special case of the more general result in the inverse problem framework, therefore all our comments below, except the identifiability issue, apply to such a regression framework as well.

3.3. Commenting the choice of the penalty

Theorem 1 suggests to take the penalty function, at least for models m for which the quantity Q_m is very large, such that $\text{pen}(m) \ge Q_m$ and choose the L_m's in such a way to get the value of the quantity $\frac{2\Sigma}{\lambda}$ very small in order to achieve a reasonable value of the upper bound of the penalized quadratic risk. Hence, in order to see the performance of the proposed model selection approach, one has to choose accordingly the penalty function and its constants, and let us assume in the remainder that the solution β obeys the linear identifiability condition.

Thus, we propose to choose the penalty function for all $m \in \mathcal{M}$ such that $\text{pen}(m) = Q_m$, in other words, one puts for all $m \in \mathcal{M}$

$$\begin{aligned}\text{pen}(m) := 2tr\Big(R^T\mathcal{K}^T\Psi_m R\Big) - \theta\|\Psi_m R\|^2 + 2\big(\frac{r_m^*}{\theta} + s_m^+\big)L_m \\ + 2\|A_m\|\sqrt{L_m + h_m} + \lambda\Big(\frac{\|A_m\|}{\sqrt{L_m} + \sqrt{L_m + h_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)^2 \quad (3.6)\end{aligned}$$

Theorem 1 then guarantees that the penalized estimator $\tilde{\beta}$ obtained with such a choice of the penalty satisfies

$$\begin{aligned}(1-\theta)\mathbb{E}\Big[\|\tilde{\beta} - \beta\|^2\Big] \leq \inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2 + \beta^T X^T\Psi_m^T\Psi_m X\beta - 2\beta^T\Psi_m X\beta \\ + (1-\theta)\|\Psi_m R\|^2 + 2\big(\frac{r_m^*}{\theta} + s_m^+\big)L_m + 2\|A_m\|\sqrt{L_m + h_m} \\ + \lambda\Big(\frac{\|A_m\|}{\sqrt{L_m} + \sqrt{L_m + h_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)^2\Big\} + \frac{2\Sigma}{\lambda}\end{aligned}$$

However, it is still difficult to see in the latter formula how the performance of the proposed penalized estimator compares with the lowest value of the predictive risk over $\mathcal{M}$ which is attained only by an oracle of course. Hence, the following proposition provides an interesting strategy of choice of the constants in the penalty function (3.6) which enforces an oracle inequality.

Proposition 2. *Assume a penalty of the form (3.6). It follows that if one defines for all $m \in \mathcal{M}$ the L_m's, for some positive sequence $\ell_m, m \in \mathcal{M}$ (these shall be referred as the model weights) and for some $\alpha \in (0,1]$, as follows*

$$L_m := \frac{\|\Psi_m R\|^6\ell_m}{(\frac{2r_m^*}{\theta} + s_m^+)\|\Psi_m R\|^4 + 4\alpha^2\|A_m\|^2\big(\|\Psi_m R\|^2 + h_m\big)}$$

and defines, for some positive number ϵ the h_m's as follows

$$h_m := \frac{\|A_m\|^2}{\epsilon^2(\frac{2r_m^*}{\theta} + s_m^+)^2}\mathbb{1}_{\frac{\|A_m\|}{2\sqrt{t_m}} \geq \epsilon(\frac{2r_m^*}{\theta} + s_m^+)}$$

with

$$t_m := \frac{\|\Psi_m R\|^6\ell_m}{(\frac{2r_m^*}{\theta} + s_m^+)\|\Psi_m R\|^4 + 4\alpha^2\|A_m\|^2\|\Psi_m R\|^2}$$

and takes

$$\lambda := \frac{\sqrt{2}\Sigma^{\frac{1}{2}}}{\sup_{m\in\mathcal{M}}\Big\{\Big(\frac{\|A_m\|}{2\sqrt{L_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)\Big\}}$$

then the penalized estimator $\tilde{\beta}$ satisfies

$$\begin{aligned}\mathbb{E}\Big[\|\tilde{\beta} - \beta\|^2\Big] \leq \frac{1}{1-\theta}\inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2 + \beta^T X^T\Psi_m^T\Psi_m X\beta - 2\beta^T\Psi_m X\beta \\ \Big(1 - \theta + \ell_m + \frac{1}{\alpha}\sqrt{\ell_m}\Big)\|\Psi_m R\|^2\Big\} + \frac{2\sqrt{2}(1+\epsilon)\Sigma^{\frac{1}{2}}}{1-\theta}\Big[\sup_{m\in\mathcal{M}}\Big\{\Big(\frac{r_m^*}{\theta} + s_m^+\Big)\Big\}\Big]\end{aligned}$$

The proof of proposition 2 shall be given in the form of a commentary of theorem 1 following our choice of the penalty (3.6).
So, let us assume a penalty of the form (3.6), and choose the L_m's for some positive sequence of model weights $\ell_m, m \in \mathcal{M}$ and for some positive number α (one can assume $\alpha \in (0,1]$ w.l.g.), as follows

$$L_m := \frac{\|\Psi_m R\|^6 \ell_m}{(\frac{2r_m^*}{\theta} + s_m^+)\|\Psi_m R\|^4 + 4\alpha^2 \|A_m\|^2 \big(\|\Psi_m R\|^2 + h_m\big)}$$

and by noticing that

$$2\big(\frac{r_m^*}{\theta} + s_m^+\big) L_m \leq \ell_m \|\Psi_m R\|^2$$

and

$$2\|A_m\|\sqrt{L_m} \leq \frac{1}{\alpha}\sqrt{\ell_m}\|\Psi_m R\|^2$$

one obtains the new upper bound of the penalized risk

$$(1-\theta)\mathbb{E}\Big[\|\tilde{\beta} - \beta\|^2\Big] \leq \inf_{m\in\mathcal{M}} \Big\{\|\beta\|^2 + \beta^T X^T \Psi_m^T \Psi_m X\beta - 2\beta^T \Psi_m X\beta + \Big(1 - \theta + \ell_m + \frac{1}{\alpha}\sqrt{\ell_m}\Big)\|\Psi_m R\|^2 + \lambda\Big(\frac{\|A_m\|}{\sqrt{L_m} + \sqrt{L_m + h_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)^2\Big\} + \frac{2\Sigma}{\lambda}$$

Now, if one puts

$$\lambda := \frac{\sqrt{2}\Sigma^{\frac{1}{2}}}{\sup_{m\in\mathcal{M}}\Big\{\Big(\frac{\|A_m\|}{2\sqrt{L_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)\Big\}}$$

one finds that $\tilde{\beta}$ satisfies

$$\mathbb{E}\Big[\|\tilde{\beta} - \beta\|^2\Big] \leq \frac{1}{1-\theta}\inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2 + \beta^T X^T \Psi_m^T \Psi_m X\beta - 2\beta^T \Psi_m X\beta + \Big(1 - \theta + \ell_m + \frac{1}{\alpha}\sqrt{\ell_m}\Big)\|\Psi_m R\|^2\Big\} + \frac{2\sqrt{2}}{1-\theta}\Big[\sup_{m\in\mathcal{M}}\Big\{\Big(\frac{\|A_m\|}{2\sqrt{L_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)^2\Big\}\Sigma\Big]^{\frac{1}{2}} \tag{3.7}$$

which recalls the oracle equality (3.3) up to the per-model multiplicative quantity $\Big(1 - \theta + \ell_m + \frac{1}{\alpha}\sqrt{\ell_m}\Big)$ and the additive constant which is given by

$$\Gamma(\theta, \alpha, \mathcal{L}) := \frac{2\sqrt{2}}{1-\theta}\Big[\sup_{m\in\mathcal{M}}\Big\{\Big(\frac{\|A_m\|}{2\sqrt{L_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)^2\Big\}\Sigma\Big]^{\frac{1}{2}}$$

and which can be made universal at the price of augmenting the ℓ_m's. It is interesting to note that one can see the quantity $\Big(1 - \theta + \ell_m + \frac{1}{\alpha}\sqrt{\ell_m}\Big)\|\Psi_m R\|^2$ as the difficulty of representing the solution β by using model m, expressed

in different words, one would say that such a quantity measures somehow the complexity of a given model m with respect to the model collection $\mathcal{M}$.

Regarding now the choice of the model weights ℓ_m, one has to distinguish generally between the two cases : a small collection of models and a large collection of models. To start, if one assumes a small collection of models (say, a few dozens), then one may take them to be fix, i.e., $\ell_m := \ell$ for all $m \in \mathcal{M}$ while guaranteeing that for a reasonable choice of the constant ℓ, the value of Σ can be bounded by a small enough constant C. In this case, one obtains over the set $\mathscr{L}$ of all collections of estimators such that $\mathscr{L} = \left\{\mathcal{L}',\ \text{s.t. } \Gamma(\theta, \alpha, \mathcal{L}') \leq C\right\}$ a near-oracle performance of the form

$$\mathbb{E}\Big[\|\tilde{\beta} - \beta\|^2\Big] \leq K \inf_{m\in\mathcal{M}} \mathbb{E}\Big[\|\hat{\beta}_m - \beta\|^2\Big] + C$$

with $K = K(\theta, \alpha, C) = \frac{1-\theta+\ell(C)+\frac{1}{\alpha}\sqrt{\ell(C)}}{1-\theta}$.

Now, assume a large family of models, then taking the ℓ_m's as a fixed value for all $m \in \mathcal{M}$ may result in an inflation of the upper bound of the penalized risk. To account for this situation, let us first define the following quantity for all $m \in \mathcal{M}$

$$\Delta_m := \frac{\|\Psi_m R\|^6 \ell_m}{(\frac{2r_m^*}{\theta} + s_m^+)\|\Psi_m R\|^4 + 4\alpha^2\|A_m\|^2(\|\Psi_m R\|^2 + h_m)}$$

which can be seen, in some sense, as a measure of a model's complexity, hence one can write $L_m = \Delta_m \ell_m$. To start in exhibiting the intuition for the subsequent choice of the ℓ_m's, let us fix ideas by assuming, for the moment, that Δ_m can take its values in an ordered set of discrete values, and for simplicity, one can assume that such values are the integers in the interval $\big[\Delta_{min}, \Delta_{max}\big]$. Let us also denote by $\text{card}_{\mathcal{M}}(\Delta) = \sum_{m\in\mathcal{M}} \mathbb{1}_{\Delta_m=\Delta}$, and put $L_m := L_{\Delta_m}$, for all $m \in \mathcal{M}$ which makes sense since one would like to roughly privilege among models with the same complexity the one(s) which make(s) a difference (i.e., achieves the lowest value) in the bias term of the penalized risk. Hence one finds that

$$\begin{aligned}\Sigma &= \sum_{m\in\mathcal{M}} \exp\big[-\ell_m\Delta_m\big] = \sum_{\Delta=\Delta_{min}}^{\Delta_{max}} \text{card}_{\mathcal{M}}(\Delta)\exp\big[-\ell_\Delta\Delta\big] \\ &= \sum_{\Delta=\Delta_{min}}^{\Delta_{max}} \exp\Big[\log\big[\text{card}_{\mathcal{M}}(\Delta)\big] - \ell_\Delta\Delta\Big]\end{aligned}$$

then, if one takes $\ell_\Delta = \delta + \frac{\log\left[\text{card}_{\mathcal{M}}(\Delta)\right]}{\Delta}$, interestingly, one finds that $\Sigma \leq \frac{\exp\left[-\Delta_{min}\delta\right]}{1-\exp[-\delta]}$; which means that with such a choice of the model weights, one can make arbitrarily small the value of Σ.

Therefore, one would like to generalize this idea to the actual case in which the Δ_m's do not take discrete values but continuous values instead in the hope of bounding sharply the value of Σ while keeping the expression of the penalized

risk reasonable. Then one way to achieve such a goal would be by mimicking the discrete quantity $\mathrm{card}_{\mathcal{M}}(\Delta_m)$ in $\mathbf{R}^+$ by the means of a kernel diffusion for instance[8]. By choosing the latter to be a gaussian kernel with a bandwidth σ, then one possible expression that mimics interestingly $\mathrm{card}_{\mathcal{M}}(\Delta_m)$ in $\mathbf{R}^+$ is

$$\sum_{m' \in \mathcal{M}} \exp\Big[- \frac{\big(\Delta_{m'} - \Delta_m\big)^2}{2\sigma^2} \Big]$$

One checks that in the discrete case, the latter quantity converges to $\mathrm{card}_{\mathcal{M}}(\Delta_m)$ as $\sigma^2 \to 0$. One may then take the ℓ_m's for some positive number δ as follows

$$\ell_m := \delta + \frac{\log\Big[\sum_{m' \in \mathcal{M}} \exp\big[- \frac{\left(\Delta_{m'} - \Delta_m\right)^2}{2\sigma^2}\big]\Big]}{\Delta_m}$$

in which case, one finds that

$$\Sigma = \sum_{m \in \mathcal{M}} \frac{\exp\big[- \delta \Delta_m \big]}{\sum_{m' \in \mathcal{M}} \exp\Big[- \frac{\left(\Delta_{m'} - \Delta_m\right)^2}{2\sigma^2}\Big]}$$

Though it is not generally possible to give a sharp closed-form expression of Σ, one has Σ which is overwhelmingly bounded by $\sum_{m \in \mathcal{M}} \exp\big[- \delta \Delta_m \big]$ for a reasonable choice of the constant σ^2 (i.e., not taking it too small), hence if one fixes the value of σ^2, a proper setting of the constant δ allows to bring down enough the value of $\Gamma(\theta, \alpha, \mathcal{L})$, i.e., in such a way to achieve $\Gamma(\theta, \alpha, \mathcal{L}) \leq C$ for some small enough constant C. One should choose the constant C to be small enough so as not to inflate the additive constant $\Gamma(\theta, \alpha, \mathcal{L})$ in the formula of the upper bound of the penalized risk, and big enough in order to keep the model weights $\ell_m, m \in \mathcal{M}$ reasonable. In practical applications, one should take it as follows $C := o(\|\beta\|^2)$.

Concerning the choice of σ, there is not an optimal value of it actually since as σ increases, the ℓ_m's decrease (desired) resulting in the increase of Σ (undesired), and vice-versa. Nevertheless, there exists some intuitive facts that suggest to take σ as follows

$$\sigma := \frac{\tau}{\mu p}$$

with μ which can be taken for example as follows $\mu := 3$, and τ which is some positive quantity which represents in some sense the scale of the quantities $\Delta_m, m \in \mathcal{M}$, which, therefore, may be taken as follows

$$\tau := \sqrt{\frac{1}{|\mathcal{M}|} \sum_{m, m' \in \mathcal{M}} \Big(\Delta_m - \Delta_{m'}\Big)^2}$$

In fact, the intuition for such a choice of σ is that, if one had to redefine discrete values for Δ_m, one would achieve this as follows $\Delta_m := [\frac{\Delta_m}{\tau}] \cdot \tau, \forall m \in \mathcal{M}$ with

[8] This is remindful of Kernel Density Estimation (KDE), the difference here is that the Δ_m's are deterministic values ...

$[x]$ which stands for the integer value of x. Hence by mapping the $|\mathcal{M}|$ values of Δ_m into the discrete interval $[1, \cdots, p]$, one realizes in some way a discretization of the Δ_m's with a discretization step which is equal to $\frac{\tau}{p}$, hence one retrieves somehow the discrete case that we discussed above. It follows that imitating the quantity $\mathrm{card}_{\mathcal{M}}(\Delta)$ amounts to taking $3\sigma \approx \frac{\tau}{p}$ (since almost 90% of the energy of a gaussian distribution with standard deviation σ is concentrated in an interval of size $\mu\sigma$ around its mean, with $\mu \approx 3$).

Regarding the setting of the h_m's, following our choice above of the penalty and some of its constants namely the L_m's and λ, their role principally is to prevent the inflation of the additive constant $\Gamma(\theta, \alpha, \mathcal{L})$ thereby the inflation of the upper bound of the penalized risk when there are among the collection $\mathcal{M}$ some models with too small weights (which can be seen as nuisance models in some sense). So, to address the choice h_m's, first let us denote by

$$t_m := \frac{\|\Psi_m R\|^6 \ell_m}{(\frac{2r_m^*}{\theta} + s_m^+)\|\Psi_m R\|^4 + 4\alpha^2 \|A_m\|^2 \|\Psi_m R\|^2}$$

hence, one may define the h_m's as follows

$$h_m := \frac{\|A_m\|^2}{\epsilon^2(\frac{2r_m^*}{\theta} + s_m^+)^2} \mathbb{I}_{\frac{\|A_m\|}{2\sqrt{t_m}} \geq \epsilon(\frac{2r_m^*}{\theta} + s_m^+)}$$

and one has

$$\frac{\|A_m\|}{\sqrt{L_m} + \sqrt{L_m + h_m}} + \frac{r_m^*}{\theta} + s_m^+ \leq (1+\epsilon)\big(\frac{r_m^*}{\theta} + s_m^+\big)$$

one finds that

$$\begin{aligned} \Gamma(\theta, \alpha, \mathcal{L}) &= \frac{2\sqrt{2}(1+\epsilon)}{1-\theta}\Big[\sup_{m\in\mathcal{M}}\Big\{\Big(\frac{r_m^*}{\theta} + s_m^+\Big)^2\Big\}\Sigma\Big]^{\frac{1}{2}} \\ &= \frac{2\sqrt{2}(1+\epsilon)\Sigma^{\frac{1}{2}}}{1-\theta}\Big[\sup_{m\in\mathcal{M}}\Big\{\Big(\frac{r_m^*}{\theta} + s_m^+\Big)\Big\}\Big] \end{aligned}$$

One may choose the constant ϵ for instance as follows $\epsilon := 1$; which turns to be a good trade-off between keeping the constant $\Gamma(\theta, \alpha, \mathcal{L})$ reasonable on the one hand, and not inflating too much the weights ℓ_m on the other hand.

For the choice of two remaining penalty parameters namely θ and α, as one could notice, their optimal choice strictly speaking doesn't exist as this suggests knowledge upfront of the best model. Nevertheless, it is easy to choose them very reasonably without prior knowledge of the best model by examining the formula of the upper bound of the penalized quadratic risk (3.7). Indeed, firstly for θ, formula (3.7) suggests to use values which lie between but far enough from the critical values 0 and 1, hence, unless there is a possibility of optimizing the value of θ for a specific application (e.g. by simulation), we suggest to choose it for example as follows $\theta := \frac{3}{4}$ which happens to be a good compromise between the

bias and the variance term in the formula of the upper bound of the penalized risk. Secondly for α, one can see that increasing α results in the decrease of the quantity $\left(1-\theta+\ell_m+\frac{1}{\alpha}\sqrt{\ell_m}\right)$ and the increase of Σ at the same time. And by remarking that the dominant quantity in $\ell_m+\frac{1}{\alpha}\sqrt{\ell_m}$ is ℓ_m, hence any value greater enough than 0 would be adequate for α, hence by default, one may take it as follows $\alpha := 1$ which happens to be a reasonable choice.

The strategy of choice of the penalty and its constants that we presented in this section constitutes of course one possible strategy that turns to enforce an oracle inequality of the selected estimator independently of any application, and we do not exclude that more optimized penalties for some specific applications– by the means for instance of a simulation study that attempts to learn the optimal values of the involved constants in the penalty for these applications– could be designed. Other papers dedicated more to applications will follow this one anyway in which the dear reader can find more hints for optimizing the penalty for some specific applications.

4. Linear identifiability assumption

Whenever the estimation of the vector β from (1.1) is the statistician's main concern (i.e., the inverse problem framework), and the rank of matrix X is inferior than p, one is faced unavoidably to an identifiability problem of model (1.1) that needs to be addressed before using the proposed model selection approach. To put forward such an issue, let us consider the following noiseless linear model

$$y' = X\beta \tag{4.1}$$

and consider the two subspaces of $\mathbf{R}^P$: $\mathcal{S}_1 = \left\{X^\dagger X t, t \in \mathbf{R}^p\right\}$, and $\mathcal{S}_2 = \left\{(I_p - X^\dagger X)t, t \in \mathbf{R}^p\right\}$. Please note that $\mathcal{S}_1 \cup \mathcal{S}_2 = \mathbf{R}^P$, and $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, one deduces that all solutions of the form

$$\beta(\eta) = X^\dagger y' + (I_p - X^\dagger X)\eta$$

for any $p-$dimensional real vector η, satisfy equation (4.1). This simply means that, unless the component of β that belongs to $\mathcal{S}_2$, i.e., $(I_p - X^\dagger X)\beta$ is known or identifiable with respect to the identifiable component $X^\dagger X\beta$, one cannot generally recover β from an instance of (4.1), hence its estimation from an observation of model (1.1) is highly problematic.

Therefore, in order to guarantee identifiability of model (1.1), the hypothesis which we put forward in section 4 and which we called the linear identifiability condition consists in saying that one knows *a priori* a linear operator $\mathcal{K}$ that we called a noiseless reconstructor of β such that one can write $\beta = \mathcal{K}X\beta$ or equivalently $(I_p - \mathcal{K}X)\beta = 0$. Such a linear identifiability condition happens in fact to be realistic for various practical applications and some of them are discussed below.

Firstly, in many engineering fields, one commonly assumes that the solution β that one would like to recover from an observation of (1.1) is compressible in

some basis Λ of $\mathbf{R}^p$ (for instance some wavelet basis (22; 27; 17)); which means that if one computed the coefficients of the scalar product between β and the respective vectors of the basis Λ, many of these coefficients would be found to be zero (or almost zero). We show indeed in this case that under some mild assumptions which happen to be realistic for numerous practical applications, one may derive easily a possible expression of the noiseless reconstructor $\mathcal{K}$ that allows to overcome the identifiability problem of the model. In short, saying that β is compressible in some basis Λ means that one is capable of extracting from Λ a subset of vectors which scalar product with β is fatally zero [9]. Hence one proceeds by extracting from Λ a number $k = p - \mathrm{rank}(X)$ of such vectors, then use them to construct the rows of a matrix ϕ so as to enforce on the solution β an equality of the form $\phi\beta = 0$. One checks easily that if matrix ϕ satisfies simultaneously the two following conditions:

1. $\phi\beta = 0$
2. the rank of the augmented matrix $\begin{bmatrix} X \\ \phi \end{bmatrix}$ –defined by the union of the rows of X and the rows of ϕ– is equal to p

then holds necessarily the following linear relationship between the two vectors β and $X\beta$:

$$\beta = \left(X^T X + \phi^T \phi\right)^{-1} X^T X\beta$$

Hence, one can define the linear reconstructor $\mathcal{K}$ of β as follows

$$\mathcal{K} = \mathcal{K}(X, \phi) := \left(X^T X + \phi^T \phi\right)^{-1} X^T$$

More generally, one shows that for any matrix Π such that the rank of the augmented matrix $\begin{bmatrix} X \\ \Pi \end{bmatrix}$ –which rows are the union of the rows of X and the rows of Π– is equal to p then one has the following linear relationship which holds for the two vectors β and $X\beta$:

$$\beta = \left(X^T X + \Pi^T \Pi\right)^{-1} X^T X\beta + \left(\Pi\big(I_p - X^\dagger X\big)\right)^\dagger \Pi\beta$$

In particular, if one chooses matrix Π in such a way to have $\|\Pi\beta\| \ll \|\beta\|$, thereby $\|\left(\Pi\big(I_p - X^\dagger X\big)\right)^\dagger \Pi\beta\| \ll \|\beta\|$, then one has

$$\beta \approx \left(X^T X + \Pi^T \Pi\right)^{-1} X^T X\beta$$

Such an idea might be useful when one a-priori knows that the solution β belongs to some subspace of $\mathbf{R}^p$ which is given by $\{t \in \mathbf{R}^p, Gt \approx t\}$ for instance a subspace of the form $\{t \in \mathbf{R}^p, (G - I_p)t \leq C\|t\|^2\}$ with $C = o(1)$ and G

[9] When for instance β is known to be smooth, any trigonometric or wavelet basis might do the job.

standing for some linear operator which is a p by p matrix different enough than the p by p identity matrix; in the sense that it can compensate for the rank deficiency of matrix X in the sense that we mentioned above. As an illustrative example, assume that the solution is smooth, then it is known that multiplying it on the left by a smoothing kernel G_σ belonging to some family of parametric kernels $\mathcal{G} = \{G_\sigma, \sigma \in \mathcal{S}\}$ (e.g. gaussian kernels with increasing bandwidth σ) would yield approximately the same solution provided evidently that the parameter σ of G_σ is set in such a way not to flatten too much the solution. In this case, one would take G_σ as the kernel in $\mathcal{G}$ with the smallest possible smoothing power corresponding, say, to a parameter σ^* and such that the augmented matrix $\begin{bmatrix} X \\ I_p - G_{\sigma^*} \end{bmatrix}$ is of full rank (i.e., p). One could then meet partially the identifiability condition of model (1.1) with matrix ϕ being equal to $I_p - G_{\sigma^*}$.

We mentioned in section 3 that one can recover, by using the proposed model selection approach, the least-norm solution $\mu = X^\dagger y' = X^\dagger X\beta$ of the equation $y' = X\beta$, in other words the solution of the optimization problem

$$\|\mu\|^2 \rightarrow \min_{X\mu = X\beta}$$

since one has $y = X\mu + Rz$ and the linear identifiability condition which is met for μ since one has $\mu = X^\dagger X\mu$. One can actually generalize such an idea in order to estimate all identifiable solutions of the form

$$\mu^T \Pi \mu \rightarrow \quad \min_{\substack{X\mu = X\beta \\ \phi\mu = 0}}$$

for some p by p positive semi-definite symmetric matrix Π and some k by p matrix ϕ ($k \leq p - \text{rank}(X)$). Then, one checks easily that, if matrix $\left(\Pi + X^T X + \phi^T \phi\right)$ is of full rank, such a solution μ is unique (i.e., identifiable) and it is given by the formula

$$\mu = \mathcal{K} X\beta = \mathcal{K} X\mu$$

where

$$\mathcal{K} = B\left(X^T - X^T X A X^T - \phi^T \phi A X^T\right) + A X^T \tag{4.2}$$

with

$$A = \left(\Pi + X^T X + \phi^T \phi\right)^{-1}$$

and

$$B = \left(X^T X + \phi^T \phi\right)^\dagger$$

In practice, one may approximate formula (4.2) for a great enough positive number μ as follows

$$\mathcal{K} \approx \mu\Big(\Pi + \mu X^T X + \mu \phi^T \phi\Big)^{-1} X^T$$

Such a notion might be useful when β belongs for instance to some ℓ_2-body (ellipsoid), in other words, one knows some orthonormal matrix Φ and a positive semi-definite matrix C such that

$$\big(\Phi\beta\big)^T C(\Phi\beta\big) \leq 1$$

Hence, one might define for example matrix Π as follows

$$\Pi := \Phi^T C \Phi$$

and matrix ϕ might be useful if one knows that some of the components of the vector of the coefficients $\Phi\beta$ are fatally zero, in this case, one constructs ϕ as follows

$$\phi = D\{\delta_k\}_{k=1,p}\Phi$$

with $\delta_k = \mathbb{1}_{(\Phi\beta)_k=0}$. A common example is when the coefficients of β with respect to some orthonormal basis Φ decay rapidly typically like a power law, which is the case of most useful practical solutions when the orthonormal basis Φ is appropriately chosen.
We would like to add that the author is currently investigating other identifiability schemes that might broaden the scope of application of the proposed method.

5. A numerical study

Please, note that our experiments below were realized with the penalty (3.6), and its constants were chosen as explained in subsection 3.3. We will show the performance of our method on two applications by using the classical performance ratio $\rho = \frac{\mathbb{E}\left[\|\tilde{\beta}-\beta\|^2\right]}{\min_{m\in\mathcal{M}}\mathbb{E}\left[\|\hat{\beta}_m-\beta\|^2\right]}$, where $\mathbb{E}\big[\|\tilde{\beta}-\beta\|^2\big]$ is estimated by averaging the value of $\|\tilde{\beta}-\beta\|^2$ over 50 instances of the noisy signal y, and $\min_{m\in\mathcal{M}}\mathbb{E}\big[\|\hat{\beta}_m-\beta\|^2\big]$ is computed directly by using formula (3.2).
The first application concerns gaussian filtering for smooth signals, and the second application concerns statistical inversion of ill-posed linear inverse problems by using regularity (smoothness) and parsimony priors on the solution.

All our experiments below were performed on the following signal (see fig. 1) and for $p = 100$:

$$\beta(t) = \frac{1}{50}\Big(\exp[-t](t/10)^2/2+(t/10)\log(t/10+1)+25\sin(t/5)\exp[t/50]\Big); \forall t = 1,\cdots,p$$

and their Matlab code is available upon mail request to the author.

5.1. Gaussian smoothing

We assume an homoscedastic linear regression model

$$y(t) = \beta(t) + z(t);\ t = 1,\cdots,p$$

TABLE 1
Some results showing the performance ratio of the proposed method versus the oracle performance for the problem of signal gaussian smoothing; for different sizes (M) of the model collection.

Gaussian smoothing by model selection	
M	ρ
50	1.7321
100	1.6943
200	1.6135
500	1.5640
1000	1.5875

where $z(t), t = 1, \cdots, p$ are i.i.d. standard gaussian variables, and we consider the collection of M linear filters $\left\{G_{\sigma_m}, m = 1, \cdots, M\right\}$ where for all $m = 1, \cdots, M$, G_{σ_m} stands for a gaussian filter (p by p symmetric matrix) with bandwidth σ_m defined as follows

$$G_{\sigma_m}(i,j) := \frac{1}{a_i^m} \exp\Big[- \frac{(i-j)^2}{2\sigma_m^2}\Big];\ \forall i,j = 1, \cdots, p$$

where for all $m = 1, \cdots, M$ and for all $i = 1, \cdots, p$, a_i^m stands for a normalization constant with respect to row i of G_{σ_m} given by

$$a_i^m := \frac{1}{p} \sum_{j=1}^{p} \exp\Big[- \frac{(i-j)^2}{2\sigma_m^2}\Big]$$

Typically, we take the σ_m's as multiples of the value $\sigma = \frac{10}{M}$ as follows $\sigma_m = m \cdot \sigma, m = 1, \cdots, M$. The goal is then to select the gaussian filter with the best bandwidth to apply on the signal y by using the described model selection approach, and the results are summarized in table 1 above.

5.2. Statistical inversion of ill-posed linear inverse problems

We assume the following linear inverse problem

$$y = X\beta + z;\ t = 1, \cdots, p$$

where $z = z(t)_{t=1,\cdots,p}$ stands for a standard $p-$dimensional gaussian vector, and X is an ill-conditioned p by p matrix. For our simulations, we generated such a matrix X randomly such that for all $i, j = 1, \cdots, p$, one has $X(i,j)$ is an i.i.d. standard gaussian variable. To check the ill-conditioning of a randomly generated matrix X, we computed the ratio between its largest and smallest singular value $\frac{s^*}{s_*}$. For the matrix X we used, we found that $s^* = 19.9659$ and $s_* = 0.0098$, hence $\frac{s^*}{s_*} \approx 2043.7$. Now, to recover β from a a noisy observation y, we used a collection of M linear filters of the form (see section 2) :

$$\Psi_m := \Big(X^T X + \mathcal{H}_m\Big)^{-1} X^T \tag{5.1}$$

where, for all $m = 1, \cdots, M$, one has $\mathcal{H}_m$ which stands for some linear combination of first, second and third order discrete differential operators as follows

$$\mathcal{H}_m = a_m \big(\mathcal{D}^1\big)^T \mathcal{D}^1 + b_m \big(\mathcal{D}^2\big)^T \mathcal{D}^2 + c_m \big(\mathcal{D}^3\big)^T \mathcal{D}^3$$

where a_m, b_m and c_m stand for three positive numbers, and $\mathcal{D}^1$, $\mathcal{D}^2$ and $\mathcal{D}^3$ stand respectively for first, second and third order differential operators given by

$$\mathcal{D}^1(i,j) = \begin{cases} -1, & \text{if } i = j. \\ 1, & \text{if } i = j-1. \\ 0, & \text{else.} \end{cases}$$

and

$$\mathcal{D}^2 = \mathcal{D}^1\mathcal{D}^1 \ ; \mathcal{D}^3 = \mathcal{D}^1\mathcal{D}^2$$

so as to achieve different degrees of smoothness in the recovered solution.
The sequence of the triplets $\{(a_m, b_m, c_m), m \in \mathcal{M}\}$ was generated as follows. So, for all $m \equiv m(i,j,k)$ for some $i, j, k \in [0, \cdots, 9]$

$$a_m := (2^i - 1) \ ; b_m := (2^j - 1) \ ; c_m := (2^k - 1)$$

so we ended up with a collection of 1000 linear filters of the form (5.1). We repeated this experience for 50 instances of the noisy signal y in order for us to be able to estimate the performance ratio $\rho = \frac{\mathbb{E}\left[\|\tilde{\beta} - \beta\|^2\right]}{\min_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{\beta}_m - \beta\|^2\right]}$ between the penalized risk and the oracle risk; so we could estimate $\rho = 4.8936$. Clearly, such a value of ρ is about three times greater than in the experiments we showed in subsection 5.1 which, at first glance, may look like an underperformance of the presented approach. This is not actually the case and this needs to be moderated for the reasons that we review here. The first reason is that the problem we treated in this subsection is much more complicated then the previous one since matrix X is severely ill-conditioned which results in larger penalties, thereby in larger upper bounds of the penalized risk. The second reason is that we used so strong priors that an oracle was enabled to recover the original signal almost perfectly. To see this, we computed for the present experiment the relative error ratio $\frac{\min_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{\beta}_m - \beta\|^2\right]}{\|\beta\|_{\cdot}^2}$ and we found a value of order of 0.2%; which means that the oracle succeeded in recovering almost perfectly the actual signal β. We compared such oracle relative error with the one of our method, and we noticed that the latter could recover the solution with a relative error of order less than 1% which is not that bad (see figure 2 for visual assessment of the recovered solution). Please, note that if one wanted to recover the solution in the present experiment by using direct inversion of the linear model as follows $X^{-1}y$, then one's expected value the relative error ratio, i.e., $\frac{\mathbb{E}\left[\|X^{-1}y - \beta\|^2\right]}{\|\beta\|_{\cdot}^2}$ would be of the order of 610% which is of course unreasonable from a practical point of view (see also figure 2 for a visual constatation).

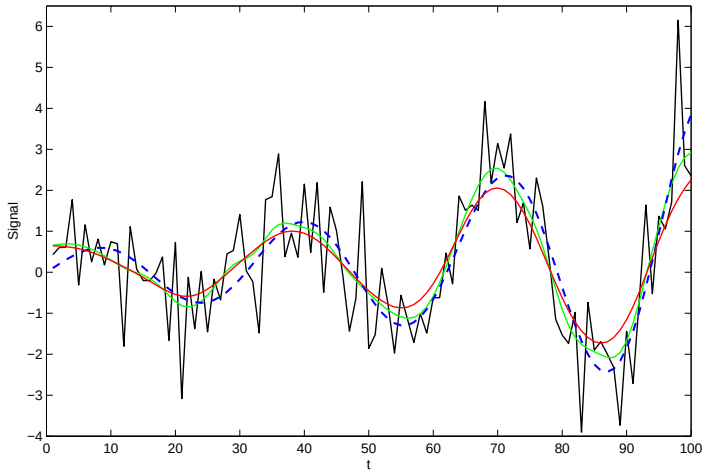

FIG 1. *An instance of gaussian smoothing of a noisy signal by using the described model selection method : the number of models is* $M = 1000$*; the oracle selected for this signal* $\sigma^* = 2.43$ *and the method selected* $\tilde{\sigma} = 4.05$. **In dashed blue***: The original signal* (β) *;* **In black***: The noisy signal (y) ;* **In green** *: The oracle-driven smoothed signal* $(\hat{\beta}^*)$ *;* **In red***: The data-driven smoothed signal* $(\tilde{\beta})$.

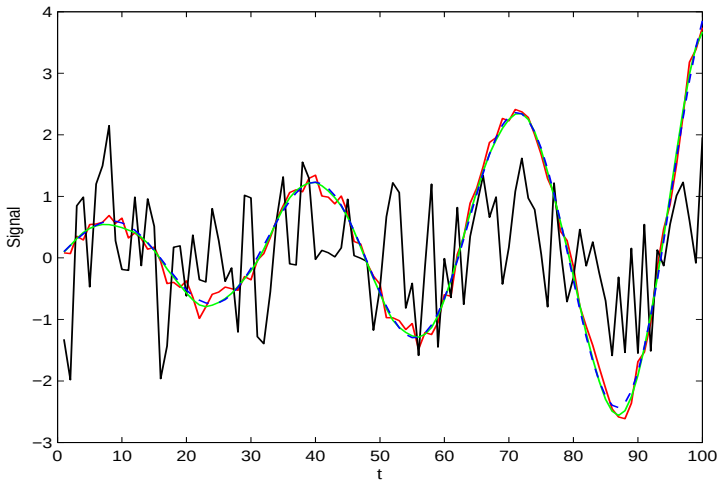

FIG 2. *An instance of a statistical inversion of a linear inverse problem with regularity prior on the solution by using the proposed model selection approach :* **In dashed blue***: The original signal* (β) *;* **In black***: The noisy signal* $(\frac{y}{3})$ *scaled by a factor of* $\frac{1}{3}$ *to allow better visualization ;* **In green** *: The oracle-driven estimated solution* $(\hat{\beta}^*)$ *;* **In red***: The data-driven estimated solution* $(\tilde{\beta})$.

6. Conclusion

We shall conclude the present paper briefly by saying that we presented a new model selection framework that addresses the problem of non least-squares linear estimation in linear regression and linear inverse problems by using model selection via penalization in the spirit of the pioneering works on non-asymptotic model selection works of Birgé and Massart mainly (12), and we showed its good performance on two practical problems renowned to be difficult to address in practical applications. Moreover, we think that with more optimized penalty constants (optimized for instance by the means of a simulation study), one could achieve even better performance of the proposed approach. We would like then to note that other papers dedicated more to applications of the proposed approach will follow the present one in which we plan to study more optimized penalties for various applications pertaining to our field of expertise (mainly signal and image processing applications).

The present paper constitutes a first attempt by the author to give a satisfactory answer to the question of non-least squares estimation in regression and inverse problems by model selection. There are, of course, still many challenges that we would like to address along with the statistical community in the near future before achieving a complete final package of the present approach, among which we review some below in the form of question/answer:

- Could our results in this paper be improved ? Our answer to this question would be by "yes". Indeed, though we personally think that the concentration inequality that we made use of to prove the main theorem in this paper is sharp (see lemma 7.1), nevertheless, if one could propose sharper concentration inequalities that could lead to smaller penalties, one might improve on our model selection results in this paper.
- What happens if one used instead smaller penalties than the one we proposed in this paper? In fact, one notices that theorem 1 does not forbid the use of smaller penalties than penalty (3.6) which we showed to enforce an oracle inequality. However, since theorem 1 only gives an upper bound of the penalized risk, hence it is difficult to figure out actually what would be the behavior of the penalized estimator if one used smaller penalties mainly for models with a value of Q_m which is too big. This is an open question that we will endeavor to solve in the near future hopefully in the spirit of the findings of Birgé & Massart in (12).
- When a linear model is underdetermined, is the linear identifiability assumption really mandatory in order for the model selection procedure to work? A rapid answer to this question would be by "yes", because such an identifiability assumption serves in some sense as a compass for the model selection procedure to go in the right direction in an attempt to recover the *sought* solution of interest (among an infinite number of candidate solutions), which turns to be rather subjective because it is imposed by the user. Nevertheless, we do not exclude that when some generic assumptions can be made on the solution of interest (by assuming for instance

that it belongs to some restrictive class of $p-$dimensional vectors), one might devise for example some (implicit) identifiability schemes that can achieve comparable results to those in this paper.

- When noise matrix R is unknown, how one should modify the penalty function to allow simultaneous estimation of β and R? We believe that this could be done in the near future for example in the spirit of the recent work of Baraud and collaborators in (5).
- What is the convergence rate of the proposed estimation approach with respect to some classes of the vector β (e.g. a Sobolev body)? We did not answer yet this question in the present paper, however, we are currently investing a significant amount of our time trying to answer this question and many other theoretical questions which could be of significant interest either from a theoretical or from a practical point of view.

By saying this, we concluded then this paper.

7. Appendix section

Appendix 1 : Proof of the main theorem

Proof. We shall use sometimes in the proof the fact that $2ab \leq \eta a^2 + \frac{b^2}{\eta}, \forall a, b, \eta > 0$.
Let us fix some $m \in \mathcal{M}$. One has one the one hand

$$\|\hat{\beta}_m - \beta\|^2 = \beta^T \big(\Psi_m X - I_p\big)^T \big(\Psi_m X - I_p\big)\beta + z^T R^T \Psi_m^T \Psi_m R z + 2\beta^T \big(\Psi_m X - I\big)^T \Psi_m R z$$

and by using the fact that $\beta = \mathcal{K} X \beta$, one finds that

$$\|\hat{\beta}_m - \beta\|^2 = \|\beta\|^2 + (X\beta)^T \Big(\Psi_m^T \Psi_m - 2\mathcal{K}^T \Psi_m\Big) X\beta + z^T R^T \Psi_m^T \Psi_m R z + 2(X\beta)^T \big(\Psi_m - \mathcal{K}\big)^T \Psi_m R z$$

On the other hand, one has

$$\mathrm{Crit}(m) = y^T \Big(\Psi_m^T \Psi_m - 2\mathcal{K}^T \Psi_m) y + \mathrm{pen}(m)$$

hence

$$\mathrm{Crit}(m) = (X\beta)^T \Big(\Psi_m^T \Psi_m - 2\mathcal{K}^T \Psi_m\Big) X\beta + z^T R^T \Big(\Psi_m^T \Psi_m - 2\mathcal{K}^T \Psi_m) R z + 2(X\beta)^T \Big(\Psi_m^T \Psi_m - \mathcal{K}^T \Psi_m - \Psi_m^T \mathcal{K}\Big) R z + \mathrm{pen}(m)$$

One derives that

$$\|\hat{\beta}_m - \beta\|^2 - \mathrm{Crit}(m) = \|\beta\|^2 + 2z^T R^T \mathcal{K}^T \Psi_m R z + 2(X\beta)^T \Psi_m^T \mathcal{K} R z - \mathrm{pen}(m)$$

Now, let us fix some $\theta \in (0,1)$, one finds that

$$(1-\theta)\|\hat{\beta}_m-\beta\|^2-\mathrm{Crit}(m) = -\theta\beta^T\big(\Psi_m X-I_p\big)^T\big(\Psi_m X-I_p\big)\beta+z^TR^T\Big(2\mathcal{K}^T\Psi_m-\theta\Psi_m^T\Psi_m\Big)Rz$$
$$+2(X\beta)^T\Big(\Psi_m^T\mathcal{K}-\theta\Psi_m^T\Psi_m+\theta\mathcal{K}^T\Psi_m\Big)Rz+\|\beta\|^2-\mathrm{pen}(m)$$

hence

$$(1-\theta)\|\hat{\beta}_m-\beta\|^2-\mathrm{Crit}(m) = -\theta\beta^T\Big(\Psi_m X-I_p\Big)^T\Big(\Psi_m X-I_p\Big)\beta+z^TR^T\Big(2\mathcal{K}^T\Psi_m-\theta\Psi_m^T\Psi_m\Big)Rz$$
$$+2\beta^T\Big(X^T\Psi_m^T\mathcal{K}-\theta X^T\Psi_m^T\Psi_m+\theta\Psi_m\Big)Rz+\|\beta\|^2-\mathrm{pen}(m)$$

and by adding and subtracting the quantity $2\beta^T\mathcal{K}Rz$ in the left side of the latter equality, one finds that

$$(1-\theta)\|\hat{\beta}_m-\beta\|^2-\mathrm{Crit}(m) = -\theta\beta^T\Big(\Psi_m X-I_p\Big)^T\Big(\Psi_m X-I_p\Big)\beta+z^TR^T\Big(2\mathcal{K}^T\Psi_m-\theta\Psi_m^T\Psi_m\Big)Rz$$
$$-2\beta^T\Big(\Psi_m X-I_p\Big)^T\Big(\theta\Psi_m-\mathcal{K}\Big)Rz+\|\beta\|^2+2\beta^T\mathcal{K}Rz-\mathrm{pen}(m)$$

If now one puts $\eta_m = \big(\Psi_m X - I_p\big)\beta$ for all $m \in \mathcal{M}$, one derives

$$(1-\theta)\|\hat{\beta}_m-\beta\|^2-\mathrm{Crit}(m) = -\theta\|\eta_m\|^2+z^TR^T\Big(2\mathcal{K}^T\Psi_m-\theta\Psi_m^T\Psi_m\Big)Rz$$
$$-2\eta_m^T\Big(\theta\Psi_m-\mathcal{K}\Big)Rz+\|\beta\|^2+2\beta^T\mathcal{K}Rz-\mathrm{pen}(m)$$

and by definition of $\tilde{\beta} = \hat{\beta}_{\tilde{m}}$, one finds that

$$(1-\theta)\|\tilde{\beta}-\beta\|^2 = -\theta\|\eta_{\tilde{m}}\|^2+z^TR^T\Big(-\theta\Psi_{\tilde{m}}^T\Psi_{\tilde{m}}+2\mathcal{K}^T\Psi_m\Big)Rz$$
$$-2\eta_{\tilde{m}}^T\Big(\theta\Psi_{\tilde{m}}-\mathcal{K}\Big)Rz-\mathrm{pen}(\tilde{m})$$
$$+\inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2+y^T\Big(\Psi_m^T\Psi_m-2\mathcal{K}^T\Psi_m)y+2\beta^T\mathcal{K}Rz+\mathrm{pen}(m)\Big\}$$

Since $\tilde{m}$ is random and can be any $m \in \mathcal{M}$, it follows that in order to control $\|\tilde{\beta}-\beta\|^2$, one needs to control uniformly, i.e., for all $m \in \mathcal{M}$ simultaneously, the expression

$$\Gamma_m = z^TR^T\Big(-\theta\Psi_m^T\Psi_m+2\mathcal{K}^T\Psi_m\Big)Rz-2\eta_m^T\Big(\theta\Psi_m-\mathcal{K}\Big)Rz$$

To this end, we shall make use of concentration inequality (7.1) of lemma (7.1). So, let us put

$$\mathcal{A}_m = R^T\Big(-\theta\Psi_m^T\Psi_m+\mathcal{K}^T\Psi_m+\Psi_m^T\mathcal{K}\Big)R$$

$$\mathcal{B}_m = \Big(\theta\Psi_m-\mathcal{K}\Big)RR^T\Big(\theta\Psi_m-\mathcal{K}\Big)^T$$

and denote by s_m^+ the largest positive eigen value of the symmetric matrix $\mathcal{A}_m$, and r_m^* the largest singular value of matrix $\mathcal{B}_m$. Then, by using concentration inequality (7.1) of lemma (7.1), one has for all $m \in \mathcal{M}$, with a probability larger than $1-\exp[-x_m]$, with $x_m \geq 0$ for all $m \in \mathcal{M}$, that

$$\begin{aligned} Y_m &\leq tr\big(A_m\big) + 2\sqrt{\|A_m\|^2 + 2\eta_m^T B_m \eta_m}\sqrt{x_m} + 2s_m^+ x_m \\ &\leq tr\big(A_m\big) + 2\|A_m\|\sqrt{x_m} + 2\sqrt{2}\sqrt{\eta_m^T B_m \eta_m}\sqrt{x_m} + 2s_m^+ x_m \\ &\leq tr\big(A_m\big) + 2\|A_m\|\sqrt{x_m} + \gamma_m \eta_m^T B_m \eta_m + \frac{2}{\gamma_m} x_m + 2s_m^+ x_m \\ &\leq tr\big(A_m\big) + 2\|A_m\|\sqrt{x_m} + \gamma_m r_m^* \|\eta_m\|^2 + 2\big(\frac{1}{\gamma_m} + s_m^+\big) x_m \end{aligned}$$

for every positive number γ_m. If one takes $r_m^* \gamma_m = \theta$, hence $\frac{1}{\gamma_m} = \frac{r_m^*}{\theta}$, one then finds that, simultaneously for all $m \in \mathcal{M}$, with a probability larger than $1 - \sum_{m\in\mathcal{M}} \exp[-x_m]$ that

$$\begin{aligned} (1-\theta)\|\tilde{\beta} - \beta\|^2 \leq tr\big(A_{\tilde{m}}\big) + 2\|A_{\tilde{m}}\|\sqrt{x_{\tilde{m}}} + 2\big(\frac{r_{\tilde{m}}^*}{\theta} + s_{\tilde{m}}^+\big) x_{\tilde{m}} - \mathrm{pen}(\tilde{m}) \\ + \inf_{m\in\mathcal{M}} \Big\{ \|\beta\|^2 + y^T \Big(\Psi_m^T \Psi_m - 2\mathcal{K}^T \Psi_m \big) y + 2\beta^T \mathcal{K} R z + \mathrm{pen}(m) \Big) \Big\} \end{aligned}$$

and by putting for an arbitrary positive number ξ: $x_m = L_m + \xi$ for all $m \in \mathcal{M}$, one derives that, simultaneously for all $m \in \mathcal{M}$, with a probability larger than $1 - \Sigma \exp[-\xi]$ that

$$\begin{aligned} (1-\theta)\|\tilde{\beta} - \beta\|^2 \leq tr\big(A_{\tilde{m}}\big) + 2\big(\frac{r_{\tilde{m}}^*}{\theta} + s_{\tilde{m}}^+\big)\Delta_{\tilde{m}} + 2\|A_{\tilde{m}}\|\sqrt{\Delta_{\tilde{m}} + \xi} + 2\big(\frac{r_{\tilde{m}}^*}{\theta} + s_{\tilde{m}}^+\big)\xi - \mathrm{pen}(\tilde{m}) \\ + \inf_{m\in\mathcal{M}} \Big\{ \|\beta\|^2 + y^T \Big(\Psi_m^T \Psi_m - 2\mathcal{K}^T \Psi_m \big) y + 2\beta^T \mathcal{K} R z + \mathrm{pen}(m) \Big) \Big\} \end{aligned}$$

We need now to separate ξ from one particular m by deriving for all $m \in \mathcal{M}$ a sharp upper bound for the expression

$$2\|A_m\|\sqrt{L_m + \xi} + 2\big(\frac{r_m^*}{\theta} + s_m^+\big)\xi$$

To do this, we consider for $m \in \mathcal{M}$ a positive number h_m, then one has $\sqrt{L_m+\xi} - \sqrt{L_m + h_m} = \frac{\xi - h_m}{\sqrt{L_m+\xi} + \sqrt{L_m+h_m}} \leq \frac{\xi}{\sqrt{L_m} + \sqrt{L_m+h_m}}$. Hence, for all $\lambda > 0$

$$\begin{aligned} 2\|A_m\|\sqrt{L_m+\xi} + 2\big(\frac{r_m^*}{\theta} + s_m^+\big)\xi &\leq 2\|A_m\|\sqrt{L_m + h_m} \\ &+ 2\Big(\frac{\|A_m\|}{\sqrt{L_m} + \sqrt{L_m+h_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)\xi \\ &\leq 2\|A_m\|\sqrt{L_m + h_m} \\ &+ \lambda\Big(\frac{\|A_m\|}{\sqrt{L_m} + \sqrt{L_m+h_m}} + \frac{r_m^*}{\theta} + s_m^+\Big)^2 + \frac{\xi^2}{\lambda} \end{aligned}$$

Now, let us put for all $m \in \mathcal{M}$

$$Q_m = tr\big(A_m\big)+2\big(\frac{r_m^*}{\theta}+s_m^+\big)L_m+2\|A_m\|\sqrt{L_m+h_m}+\lambda\Big(\frac{\|A_m\|}{\sqrt{L_m}+\sqrt{L_m+h_m}}+\frac{r_m^*}{\theta}+s_m^+\Big)^2$$

hence with a probability larger than $1-\Sigma\exp[-\xi]$ that

$$(1-\theta)\|\tilde{\beta}-\beta\|^2 \leq Q_{\tilde{m}} - \mathrm{pen}(\tilde{m}) + \frac{\xi^2}{\lambda} + \inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2 + y^T\Big(\Psi_m^T\Psi_m - 2\mathcal{K}^T\Psi_m)y + 2\beta^T\mathcal{K}Rz + \mathrm{pen}(m)\Big\}$$

and after integration over all values of ξ, one finds that

$$(1-\theta)\mathbb{E}\Big[\|\tilde{\beta}-\beta\|^2\Big] \leq \sup_{m\in\mathcal{M}}\Big\{Q_m - \mathrm{pen}(m)\Big\} + \frac{2\Sigma}{\lambda}$$
$$+\inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2+\beta^T X^T\Psi_m^T\Psi_m X\beta-2\beta^T\Psi_m X\beta+\|\Psi_m R\|^2-2tr\Big(R^T\mathcal{K}^T\Psi_m R\Big)+\mathrm{pen}(m)\Big\}$$

since

$$\mathbb{E}\Big[\inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2 + y^T\Big(\Psi_m^T\Psi_m - 2\mathcal{K}^T\Psi_m)y + 2\beta^T\mathcal{K}Rz + \mathrm{pen}(m)\Big\}\Big]$$
$$\leq \inf_{m\in\mathcal{M}}\Big\{\mathbb{E}\Big[\|\beta\|^2 + y^T\Big(\Psi_m^T\Psi_m - 2\mathcal{K}^T\Psi_m)y + 2\beta^T\mathcal{K}Rz + \mathrm{pen}(m)\Big]\Big\}$$
$$= \inf_{m\in\mathcal{M}}\Big\{\|\beta\|^2+\beta^T X^T\Psi_m^T\Psi_m X\beta-2\beta^T\Psi_m X\beta+\|\Psi_m R\|^2-2tr\Big(R^T\mathcal{K}^T\Psi_m R\Big)+\mathrm{pen}(m)\Big\}$$

□

Appendix 2 : A useful concentration inequality and its proof

Lemma 7.1. *Consider the random process: $T = z^T Az + b^T z$, where A is p by p real square matrix, b is a $p-$dimensional real vector, and $z = (z_k)_{k=1,p}$ is a $p-$dimensional standard gaussian vector, i.e., $z_k, k = 1, p$ are i.i.d. zero-mean gaussian variables with standard deviation 1. let us denote by $s_k, k = 1, p$ the respective eigen values of the symmetric matrix $\frac{1}{2}\big(A + A^T\big)$, and let us put $s^+ = \sup\{\sup_{k=1,\cdots,p}\{s_k\}, 0\}$, and $s^- = \sup\{\sup_{k=1,\cdots,p}\{-s_k\}, 0\}$. Then, the following two concentration inequalities hold true for all $x > 0$*

$$\mathbb{P}\Big[T \geq tr(A) + 2\sqrt{\frac{1}{4}\|A+A^T\|^2 + \frac{1}{2}\|b\|^2}\sqrt{x} + 2s^+x\Big] \leq \exp[-x] \qquad (7.1)$$

$$\mathbb{P}\Big[T \leq tr(A) - 2\sqrt{\frac{1}{4}\|A+A^T\|^2 + \frac{1}{2}\|b\|^2}\sqrt{x} - 2s^-x\Big] \leq \exp[-x] \qquad (7.2)$$

Proof. We already proved this lemma in (9), so we redo such a proof here. We shall make use of the result of lemma (7.2) below to prove lemma (7.1). To do this, let us consider the random process $T = z^T Az + b^T z$, that one can rewrite as follows: $T = \frac{1}{2} z^T \big(A + A^T\big) z + b^T z$, then by using the eigen value decomposition of the symmetric matrix $\frac{1}{2}\big(A + A^T\big)$, one derives $T = \sum_{k=1}^{p} s_k z'^2_k + b'_k z'$, where $s_k, k = 1, p$ stand for the respective eigen values of $\frac{1}{2}\big(A + A^T\big)$, $z' = U^T z$ with U standing for the (orthonormal) eigen matrix of $\frac{1}{2}\big(A + A^T\big)$, and $b' = U^T b$. By noticing that z' stands for a $p-$dimensional standard gaussian vector, $\|b'\|^2 = \|b\|^2$, $\sum_{k=1}^{p} s_k = tr(A)$, and $\sum_{k=1}^{p} s_k^2 = \frac{1}{4}\big\|A + A^T\big\|^2$, so by applying lemma (7.2), the proof of lemma (7.1) follows immediately. □

Lemma 7.2. *Let $a = (a_k)_{k=1,p}$ and $b = (b_k)_{k=1,p}$ be two $p-$dimensional real vectors, and consider the following random expression : $T = \sum_{k=1}^{p} a_k z_k^2 + b_k z_k$, where $z_k, k = 1, \cdots, p$ are i.i.d. $N(0,1)$, and let us put : $a^+ = \sup\{\sup_{k=1,\cdots,p}\{a_k\}, 0\}$, $a^- = \sup\{\sup_{k=1,\cdots,p}\{-a_k\}, 0\}$. Then the following two concentration inequalities hold true for all real positive x :*

$$\mathbb{P}\Big[T \geq \sum_{k=1}^{p} a_k + 2\sqrt{\sum_{k=1}^{p} a_k^2 + \frac{b_k^2}{2}}\sqrt{x} + 2a^+ x\Big] \leq \exp[-x] \tag{7.3}$$

$$\mathbb{P}\Big[T \leq \sum_{k=1}^{p} a_k - 2\sqrt{\sum_{k=1}^{p} a_k^2 + \frac{b_k^2}{2}}\sqrt{x} - 2a^- x\Big] \leq \exp[-x] \tag{7.4}$$

Proof. This lemma was also proven in (9), so we redo the proof here. We shall make use of lemma 7.3 to prove lemma (7.2). Now, to prove lemma (7.2), first, one can notice that concentration inequality (7.4) can be obtained from (7.3) by considering the random quantity

$$T' = -T = \sum_{k=1}^{p} (-a_k) z_k^2 + (-b_k) z_k$$

and by applying (7.3) on T' instead of T. So, we need to prove only (7.3). To do this, let us rewrite T as follows: $T = \sum_{k=1}^{p} T_k$, where $T_k = a_k z_k^2 + b_k z_k$, and let us compute $\log\big[\mathbb{E}\big(\exp(y(T - \bar{T}))\big)\big]$, where $\bar{T} = \sum_{k=1}^{p} a_k$. We have

$$\mathbb{E}\big[\exp(yT_k)\big] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\Big[-\frac{1}{2}((1 - 2a_k y)t^2 - 2yb_k t)\Big] dt$$

$$\mathbb{E}\big[\exp[yT_k]\big] = \exp\left[\frac{\frac{b_k^2}{2} y^2}{1 - 2a_k y}\right] \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\Big(\sqrt{1 - 2a_k y}\, t - \frac{b_k y}{\sqrt{1 - 2a_k y}}\Big)^2\right] \mathrm{d}t\right)$$

$$\mathbb{E}\big[\exp[yT_k]\big] = \frac{\exp\left[\frac{\frac{b_k^2}{2} y^2}{1 - 2a_k y}\right]}{\sqrt{1 - 2a_k y}}$$

$$\mathbb{E}\Big[\exp[y(T_k - a_k)]\Big] = \frac{\exp\left[\frac{\frac{b_k^2}{2}y^2}{1-2a_k y}\right]\exp[-ya_k]}{\sqrt{1-2a_k y}}$$

$$\log\Big(\mathbb{E}\Big[\exp[y(T_k - a_k)]\Big]\Big) = \frac{\frac{b_k^2}{2}y^2}{1-2a_k y} - \frac{1}{2}\log\Big(1-2a_k y\Big) - a_k y$$

Then, by putting $a^+ = \sup\big\{\sup_{k=1,\cdots,p}\{a_k\}, 0\big\}$, one derives (see the technical details below) that for all $0 < y < \frac{1}{2a^+}$

$$\log\Big(\mathbb{E}\Big[\exp[y(T_k - a_k)]\Big]\Big) \leq \frac{(a_k^2 + \frac{b_k^2}{2})y^2}{1-2a^+ y}$$

which implies by independence that for all $0 < y < \frac{1}{2a^+}$

$$\log\Big(\mathbb{E}\Big[\exp[y(T - \bar{T})]\Big]\Big) \leq \sum_{k=1}^{p} \frac{(a_k^2 + \frac{b_k^2}{2})y^2}{1-2a^+ y}$$

$$\log\Big(\mathbb{E}\Big[\exp[y(T - \bar{T})]\Big]\Big) \leq \frac{\Big(\sum_{k=1}^{p}\big(a_k^2 + \frac{b_k^2}{2}\big)\Big)y^2}{1-2a^+ y}$$

Finally, by applying Lemma (**??**) below with $u = \sqrt{\sum_{k=1}^{p}(a_k^2 + \frac{b_k^2}{2})}$, and $v = 2a^+$, one derives that for all $x > 0$:

$$\mathbb{P}\Big[T \geq [\sum_{k=1}^{p} a_k] + 2\sqrt{\sum_{k=1}^{p}\big(a_k^2 + \frac{b_k^2}{2}\big)}\sqrt{x} + 2a^+ x\Big] \leq \exp[-x]$$

This terminates the proof of lemma (7.2) □

Some additional technical details about the proof of lemma (7.2)

We will show in here that for all $r > 0$, $a \geq r$, and $0 < y < \frac{1}{2a}$, one has

$$\frac{-1}{2}\log(1-2ry) - ry \leq \frac{r^2 y^2}{1-2ay} \tag{7.5}$$

and that for all $r \leq 0$, for all $a > 0$, and for all $0 < y < \frac{1}{2a}$, one has

$$-\frac{1}{2}\log(1-2ry) - ry \leq \frac{r^2 y^2}{1-2ay} \tag{7.6}$$

Proof. let us start by showing inequality (7.5). To do this, let us consider the following function

$$f_{r,a}(y) = -\frac{1}{2}\log(1-2ry) - ry - \frac{r^2y^2}{1-2ay}$$

One first notices that $f_{r,a}(0) = 0$, then a sufficient condition for inequality (7.5) to hold true is that $f_{r,a}(y)' \leq 0$, for all $0 < y < \frac{1}{2a}$. We have

$$f_{r,a}(y) = \frac{-1}{2}\log(1-2ry) - ry + \frac{r^2y}{2a} + \frac{r^2}{(2a)^2} - \frac{\frac{r^2}{(2a)^2}}{1-2ay}$$

one then derives that

$$f_{r,a}(y)' = \frac{r}{1-2ry} - r + \frac{r^2}{2a} - \frac{\frac{r^2}{2a}}{(1-2ay)^2}$$

$$f_{r,a}(y)' = \frac{2r^2y}{1-2ry} - \frac{r^2y}{(1-2ay)} - \frac{r^2y}{(1-2ay)^2}$$

$$f_{r,a}(y)' \leq \frac{2r^2y}{1-2ry} - \frac{2r^2y}{(1-2ay)}$$

and finally since $\frac{1}{(1-2ry)} \leq \frac{1}{(1-2ay)}$, one deduces that

$$f_{r,a}(y)' \leq \frac{2r^2y}{(1-2ay)} - \frac{2r^2y}{(1-2ay)} = 0$$

then we have shown (7.5).
We proceed in the same way as for showing inequality (7.5) to show inequality (7.6). So let us consider the following function

$$g_{r,a}(y) = \frac{-1}{2}\log(1-2ry) - ry - \frac{r^2y^2}{1-2ay}$$

One first notices that $g_{r,a}(0) = 0$, then a sufficient condition for inequality (7.6) to hold true is to that $g_{r,a}(y)' \leq 0$ for all $0 < y < \frac{1}{2a}$. One derives that

$$g_{r,a}(y)' = \frac{r}{1-2ry} - r + \frac{r^2}{2a} - \frac{\frac{r^2}{2a}}{(1-2ay)^2}$$

$$g_{r,a}(y)' = \frac{2r^2y}{1-2ry} - \frac{r^2y}{(1-2ay)} - \frac{r^2y}{(1-2ay)^2}$$

$$g_{r,a}(y)' \leq \frac{2r^2y}{1-2ry} - \frac{2r^2y}{(1-2ay)}$$

and finally, since $\frac{1}{1-2ry} \leq \frac{1}{(1-2ay)}$, one finds that

$$g_{r,a}(y)' \leq \frac{2r^2y}{(1-2ay)} - \frac{2r^2y}{(1-2ay)} = 0$$

□

Birge's & Massart concentration inequality

Lemma 7.3. *If a random variable ξ satisfies for some two real positive numbers u and v the following inequality :*

$$\log\Big(\mathbb{E}\Big[\exp[y\xi]\Big]\Big) \leq \frac{(uy)^2}{1-vy}, \textit{for all} \quad 0 < y < \frac{1}{v} \tag{7.7}$$

then

$$\mathbb{P}\Big[\xi \geq 2u\sqrt{x} + vx\Big] \leq \exp[-x], \textit{for all} \quad x > 0 \tag{7.8}$$

The proof of this lemma can be found in (10).

Acknowledgements

This work has been accomplished by the author jointly at the Oxford Centre for Magnetic Resonance Imaging of Oxford University (U.K.), and at INRIA Nice Sophia Antipolis (France). The author is very grateful to : Alain Trubuil (INRA of Jouy-en-Josas), Christine Graffigne (MAP5, Universite Paris 5), Sylvie Huet (INRA of Jouy-en-Josas), Matthew Robson (OCMR & FMRIB, Oxford University), Alison Noble (BioMedIA, Oxford University) and Ilias Kylintireas (OCMR, Oxford University) for all the helpful discussions either about statistical model selection or image processing in video-microscopy and MRI during my two scientific stays at INRA of Jouy-en-Josas and at OCMR/FMRIB of Oxford University respectively.

References

[1] Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principal. In Proceedings 2nd International Symposium on Information Theory, P.N. Petrov and F. Csaki (Eds.).* Akademia Kiado, Budapest, 267–281.

[2] Baraud, Y. (2000). *Model selection for regression on a fixed design. Probability Theory and Related Fields.* **117** 467–493.

[3] Baraud, Y., Compte, F. and Viennet, G. (2001). *Adaptive estimation in autoregression or β-mixing regression via model selection. The Annals of Statistics.* **29** 839–875.

[4] Baraud, Y. (2002). *Model selection for regression on a random design. SAIM: Probability and Statistics* **6** 127–146.

[5] Baraud, Y. (2002). *Gaussian model selection with an unknown variance. Annals of Statistics* **37(2)** 630–672.

[6] Barron, A. (1991). *Minimum complexity density estimation. IEEE Transactions on Information Theory* **37** 1034–1054.

[7] Barron, A., Birgé, L., and Massart, P. (1999). *Risk bounds for model selection via penalization. Probab. Theory Related Fields* **113** 301–413.

[8] Bechar, I. (2009). *Non-asymptotic model selection in correlated regression. Working Draft - Available upon mail request to the author.*

[9] Bechar, I. (2009). *A concentration inequality and its application to model selection in linear inverse problems. Submitted to Acad. Sci. Paris - Available upon mail request to the author.*

[10] Birgé, L., and Massart, P. (1998). *Minimum contrast estimators on sieves: exponential bounds and rates of convergence. Bernoulli* **4** 329–375..

[11] Birgé, L., and Massart, P. (2001). *Gaussian model selection. J. Eur. Math. Soc.* **3** 203-268.

[12] Birgé, L., and Massart, P. (2007). *Minimal Penalties for Gaussian Model Selection. Probab. Theory Related Fields.* **138** 33-73.

[13] Candes, E., and Tao, T. (2007). *The Dantzig selector: statistical estimation when p is much larger than n. Annals of Statistics.* **35** 2313–2351.

[14] Candes, E., and Tao, T. (2007). *Rejoinder: The Dantzig selector: statistical estimation when p is much larger than n. Annals of Statistics.* **35(6)** 2392–2404..

[15] Candes, E., Watkin, M. and Boyd, S. (2008). *Enhancing sparsity by reweighted ℓ_1 minimization. J. Fourier Anal. Appl..* **14** 877-905.

[16] Candes, E., and Plan, Y. (2009). *Near-ideal model selection by l1 minimization. Annals of Statistics.* **37** 2145–2177

[17] Daubechies, I. (1992). *Ten Lectures on Wavelets. Number 61 in CBMS-NSF Series in Applied Mathematics.* SIAM, Philadelphia.

[18] Donoho, D.L. (2006). *Compressed sensing. IEEE Trans. Inform. Theory.* **52 no. 4** 1289-1306.

[19] Gendre, X. (2008). *Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression. Electron. J. Statist.* **2** 1345–1372.

[20] Gendre, X. (2009). *Estimation par sélection de modèle en régression hétéroscédastique. Thesis in French and English.* Universite de Nice Sophia Antipolis. url: http://math.unice.fr/ gendre/ecrits/these.pdf

[21] Johnstone, I.M. (1999). *Wavelet shrinkage for correlated data and inverse problems : Adaptivity results. Statistica Sinica* **9** 51–83.

[22] Mallat, S. (1989). *A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattn. Anal. Mach. Intell., PAMI* **11** 674–693.

[23] Mallows, C.L. (1973). *Some comments on Cp. Technometrics.* **15** 661–675.

[24] Massart, P.(2000) *Some applications of concentration inequalities to statistics. Annales de la Faculté des sciences de Toulouse, 6eme série, Tom 9.* **2** 245–303.

[25] Massart, P.(2007) *Concentration inequalities and model selection. Lecture Notes in Mathematics. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory, Saint-Flour, July 6-23, 2003.*

[26] McQuarrie, A.D.R., and Tsai, C-L. (1998). *Regression and Times Series Model Selection. World Scientific Publishing Co, Inc..* River Edge, NJ.

[27] Meyer, Y. (1990). *Ondelettes et Operateurs.* Hermann, Paris.

[28] Schwartz, G. (1978). *Estimating the dimension of a model. Ann. Statist..*

6 461–464.

[29] Shibata, R. (1981). *An optimal selection of regression variables. Biometrika* **68** 45–54.

[30] Shibata, R. (1980). *Asymptotically effcient selection of the order of the model for estimating parameters of linear process.. Ann. Statist.* **8** 147–164.

[31] Talagrand, M.(1996) *New concentration inequalities in product spaces. Invent. Math..* **126** 505–563.

[32] Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B.* **58** 267–288.

[33] Tikhonov, A.N. and Arsenin, V.A. (1977). *Solution of Ill-posed Problems. Winston & Sons*. Washington, ISBN 0-470-99124-0.